ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

LÊ QUANG HÙNG

KHAI PHÁ TRI THỨC SONG NGỮ VÀ ỨNG DỤNG TRONG DỊCH MÁY ANH – VIỆT

LUẬN ÁN TIẾN SĨ KHOA HỌC MÁY TÍNH

Hà Nội – 2016

ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

LÊ QUANG HÙNG

KHAI PHÁ TRI THỨC SONG NGỮ VÀ ỨNG DỤNG TRONG DỊCH MÁY ANH – VIỆT

Chuyên ngành: Khoa học máy tính
Mã số: 62 48 01 01

LUẬN ÁN TIẾN SĨ KHOA HỌC MÁY TÍNH

NGƯỜI HƯỚNG DẪN KHOA HỌC:

1. PGS.TS. Lê Anh Cường
2. PGS.TS. Huỳnh Văn Nam

Hà Nội – 2016

Lời cam đoan

Tôi xin cam đoan luận án này là kết quả nghiên cứu của tôi, được thực hiện dưới sự hướng dẫn của PGS.TS. Lê Anh Cường và PGS.TS. Huỳnh Văn Nam. Các nội dung trích dẫn từ các nghiên cứu của các tác giả khác mà tôi trình bày trong luận án này đã được ghi rõ nguồn trong phần tài liệu tham khảo.

Lê Quang Hùng

Tóm tắt

Nhiệm vụ của một hệ thống dịch máy là tự động dịch một văn bản từ ngôn ngữ này (ví dụ, tiếng Anh) sang một văn bản tương đương ở ngôn ngữ khác (ví dụ, tiếng Việt). Tính hữu ích của công nghệ dịch máy tăng lên cùng với chất lượng của nó. Dịch máy có nhiều ứng dụng như: (i) dịch tài liệu tiếng nước ngoài cho mục đích hiểu nội dung, (ii) dịch văn bản để xuất bản ở các ngôn ngữ khác và (iii) thông tin liên lạc, chẳng hạn như dịch *email*, *chat*, vv.

Có một số cách tiếp cận cho bài toán dịch máy như dịch trực tiếp (direct translation), dịch dựa trên chuyển đổi (transfer - based translation), dịch liên ngữ (interlingua translation), dịch dựa trên ví dụ (example - based translation) và dịch thống kê (statistical translation). Hiện tại, dịch máy dựa trên cách tiếp cận thống kê đang là một hướng phát triển đầy tiềm năng bởi những ưu điểm vượt trội so với các cách tiếp cận khác. Thay vì xây dựng các từ điển, các quy luật chuyển đổi bằng tay, dịch máy thống kê tự động xây dựng các từ điển, các quy luật dựa trên kết quả thống kê có được từ ngữ liệu. Đối với một hệ thống dịch máy thống kê, hiệu quả (chất lượng dịch) của nó tỷ lệ thuận với số lượng (kích thước) và chất lượng của ngữ liệu song ngữ được sử dụng để xây dựng hệ thống dịch. Tuy nhiên, ngữ liệu song ngữ sẵn có hiện vẫn còn hạn chế cả về kích thước lẫn chất lượng, ngay cả đối với các cặp ngôn ngữ chính. Ngoài ra, đối với các cặp ngôn ngữ có nhiều khác biệt về cấu trúc ngữ pháp (ví dụ, Anh - Việt), vấn đề về chất lượng dịch đang là thách thức đối với các nhà nghiên cứu về dịch máy trong nhiều năm qua. Vì vậy, việc bổ sung thêm ngữ liệu song ngữ và phát triển các phương pháp hiệu quả hơn dựa trên ngữ liệu hiện có là những giải pháp quan trọng để tăng chất lượng dịch cho dịch máy thống kê.

Luận án của chúng tôi tập trung giải quyết các tồn tại đã nêu thông qua ba bài toán: phát triển phương pháp xây dựng ngữ liệu song ngữ, cải tiến các phương pháp gióng hàng từ và xác định cụm từ song ngữ cho dịch máy thống kê, cụ thể như sau:

Thứ nhất, đối với bài toán xây dựng ngữ liệu song ngữ, chúng tôi khai thác từ hai nguồn: Web và sách điện tử song ngữ. Đối với nguồn từ Web, chúng tôi tập trung vào rút trích các văn bản song ngữ từ các *web-site* song ngữ. Chúng tôi đề xuất hai phương pháp thiết kế các đặc trưng dựa trên nội dung: sử dụng các từ bất biến giữa hai ngôn ngữ (cognate) và sử dụng các phân đoạn dịch. Ngoài ra,

chúng tôi kết hợp các đặc trưng dựa trên nội dung với các đặc trưng dựa trên cấu trúc của trang *web* để rút trích các văn bản song ngữ, bằng cách sử dụng phương pháp học máy. Đối với nguồn từ sách điện tử, chúng tôi đề xuất phương pháp dựa trên nội dung, sử dụng một số mẫu liên kết giữa các khối văn bản trong hai ngôn ngữ để rút trích các câu song ngữ.

Thứ hai, với bài toán gióng hàng từ, chúng tôi đề xuất một số cải tiến đối với mô hình IBM 1 theo cách tiếp cận dựa trên ràng buộc, bao gồm: ràng buộc neo, ràng buộc về vị trí của từ, ràng buộc về từ loại và ràng buộc về cụm từ. Với mỗi ràng buộc, chúng tôi đưa ra phương pháp tổng quát để tích hợp nó vào thuật toán cực đại kỳ vọng trong quá trình ước lượng tham số của mô hình. Ngoài ra, chúng tôi đưa ra một phương pháp để kết hợp các ràng buộc. Những cải tiến này đã giúp nâng cao chất lượng dịch cho hệ thống dịch máy thống kê Anh - Việt.

Thứ ba, đối với bài toán xác định cụm từ song ngữ cho dịch máy thống kê, chúng tôi đề xuất phương pháp rút trích cụm từ song ngữ từ ngữ liệu song ngữ, sử dụng các mẫu cú pháp kết hợp với gióng hàng cụm từ. Các cụm từ song ngữ này đã được ứng dụng vào việc nâng cao chất lượng dịch cho hệ thống dịch máy thống kê Anh - Việt.

Từ khóa: dịch máy, dịch máy thống kê, tri thức song ngữ, ngữ liệu song ngữ, văn bản song ngữ, gióng hàng từ.

Lời cảm ơn

Trước hết, tôi xin gửi lời cảm ơn sâu sắc đến PGS.TS. Lê Anh Cường và PGS.TS. Huỳnh Văn Nam, hai Thầy đã trực tiếp hướng dẫn, chỉ bảo tận tình, luôn hỗ trợ và tạo những điều kiện tốt nhất cho tôi học tập và nghiên cứu.

Tôi xin gửi lời cảm ơn đến các Thầy/Cô giáo ở Khoa Công nghệ thông tin, Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội, đặc biệt là PGS.TS. Phạm Bảo Sơn và các Thầy/Cô giáo ở Bộ môn Khoa học máy tính, những người đã trực tiếp giảng dạy và giúp đỡ tôi trong quá trình học tập và nghiên cứu ở trường.

Tôi xin gửi lời cảm ơn đến các đồng nghiệp ở Khoa Công nghệ thông tin, Trường Đại học Quy Nhơn, đặc biệt là TS. Trần Thiên Thành và TS. Lê Xuân Việt đã quan tâm, giúp đỡ và tạo điều kiện cho tôi trong thời gian làm nghiên cứu sinh.

Tôi xin gửi cảm ơn đến PGS.TS. Nguyễn Phương Thái, TS. Nguyễn Văn Vinh, TS. Phan Xuân Hiếu (Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội), PGS.TS. Lê Thanh Hương (Trường Đại học Bách khoa Hà Nội), TS. Nguyễn Thị Minh Huyền, TS. Lê Hồng Phương (Trường Đại học Khoa học Tự nhiên, Đại học Quốc gia Hà Nội), TS. Nguyễn Đức Dũng (Viện Công nghệ thông tin, Viện Hàn lâm Khoa học và Công nghệ Việt Nam), các Thầy/Cô đã có những góp ý chỉnh sửa để tôi hoàn thiện luận án.

Tôi xin gửi lời cảm ơn đến tất cả anh, chị, em và bạn đồng học ở Bộ môn Khoa học máy tính (Khoa Công nghệ thông tin, Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội), đặc biệt là chị Nguyễn Thị Xuân Hương (Khoa Công nghệ thông tin, Trường Đại học Dân lập Hải Phòng), nghiên cứu sinh Hoàng Thị Điệp (Khoa Công nghệ thông tin, Trường Đại học Công nghệ) đã giúp đỡ tôi trong thời gian làm nghiên cứu sinh.

Cuối cùng, tôi xin gửi lời cảm ơn đến tất cả các thành viên trong gia đình tôi, đặc biệt là vợ tôi - người đã luôn ủng hộ, chia sẽ, động viên và gánh vác công việc gia đình để tôi yên tâm học tập, nghiên cứu.

Mục lục

Danh mục các chữ viết tắt

EM	**E**xpectation **M**aximization (Cực đại kỳ vọng)
HTML	**H**yper**T**ext **M**arkup **L**anguage (Ngôn ngữ đánh dấu siêu văn bản)
ME	**M**aximum **E**ntropy (Độ hỗn loạn cực đại)
MLE	**M**aximum **L**ikelihood **E**stimation (Ước lượng khả năng cực đại)
MT	**M**achine **T**ranslation (Dịch máy)
NLP	**N**atural **L**anguage **P**rocessing (Xử lý ngôn ngữ tự nhiên)
POS	**P**art **O**f **S**peech (Nhãn từ loại)
SMT	**S**tatistical **M**achine **T**ranslation (Dịch máy thống kê)
SVM	**S**upport **V**ector **M**achine (Máy véc-tơ hỗ trợ)

Danh sách hình vẽ

Danh sách bảng

Mở đầu

1. Tính cấp thiết của luận án

Ý tưởng về dịch máy ra đời từ năm 1949 [60]. Từ đó đến nay, sau hơn 60 năm nghiên cứu và phát triển, các dịch vụ dịch máy bây giờ đã trở nên phổ biến rộng rãi. Hiện nay, có một số hệ thống dịch máy thương mại đã được sử dụng phổ biến trên thế giới như Systrans[1], Kant[2] hay những hệ thống dịch máy mở, tiêu biểu như Google[3] hỗ trợ hơn 50 cặp ngôn ngữ như Anh - Pháp, Anh - Trung, Anh - Việt, vv. Ở Việt Nam, dịch máy đã trở thành chủ đề được một số nhóm tập trung nghiên cứu. Trong số đó, có một số sản phẩm như phần mềm dịch tự động EVTRAN - một hệ thống dịch Anh - Việt hay hệ thống dịch tự động Anh – Việt của Công ty cổ phần tin học Lạc Việt[4], vv. Các cách tiếp cận cho bài toán dịch máy gồm có: dịch trực tiếp, dịch dựa trên chuyển đổi, dịch liên ngữ, dịch dựa trên ví dụ và dịch thống kê. Hiện nay, dịch máy dựa trên cách tiếp cận thống kê đang là một hướng phát triển đầy tiềm năng bởi những ưu điểm vượt trội so với các cách tiếp cận khác.

Đối với một hệ thống dịch máy thống kê, chất lượng dịch tỷ lệ thuận với số lượng và chất lượng của ngữ liệu song ngữ được sử dụng để xây dựng hệ thống dịch. Tuy nhiên, ngữ liệu song ngữ hiện vẫn còn hạn chế cả về kích thước lẫn chất lượng, ngay cả đối với các ngôn ngữ chính. Ngoài ra, đối với các cặp ngôn ngữ có nhiều khác biệt về cấu trúc ngữ pháp (ví dụ, Anh - Việt), vấn đề về chất lượng dịch đang là thách thức đối với các nhà nghiên cứu về dịch máy trong nhiều năm qua. Vì vậy, các nghiên cứu nhằm khai thác thêm ngữ liệu song ngữ và phát triển các phương pháp hiệu quả hơn dựa trên ngữ liệu hiện có để tăng chất lượng dịch cho dịch máy thống kê là những vấn đề cấp thiết và mang tính thời sự trong lĩnh vực xử lý ngôn ngữ tự nhiên hiện nay. Điều này là động lực để chúng tôi lựa chọn nghiên cứu về đề tài "Khai phá tri thức song ngữ và ứng dụng trong dịch máy Anh - Việt".

2. Mục tiêu của luận án

Trong luận án này, chúng tôi đặt ra hai mục tiêu chính:

[1]http://www.systransoft.com/lp/machine-translation/
[2]http://www.lti.cs.cmu.edu/Research/Kant/
[3]http://translate.google.com
[4]http://www.vietgle.vn/home/

- Thứ nhất, nghiên cứu đề xuất một số phương pháp để khai thác tri thức song ngữ nhằm bổ sung nguồn ngữ liệu cho dịch máy thống kê.

- Thứ hai, nghiên cứu đề xuất một số phương pháp để làm tăng chất lượng dịch cho dịch máy thống kê dựa trên ngữ liệu hiện có.

3. Đóng góp của luận án

- Đề xuất một số phương pháp để xây dựng ngữ liệu song ngữ cho dịch máy thống kê từ Web và sách điện tử song ngữ. Đối với nguồn từ Web, chúng tôi đề xuất hai phương pháp thiết kế các đặc trưng dựa trên nội dung: sử dụng *cognate* và sử dụng các phân đoạn dịch. Đối với nguồn từ sách điện tử, chúng tôi đề xuất phương pháp dựa trên nội dung, sử dụng một số mẫu liên kết giữa các khối văn bản trong hai ngôn ngữ để rút trích các câu song ngữ. Đóng góp này đã được công bố ở kỷ yếu hội thảo quốc tế *Knowledge and Systems Engineering (KSE)* năm 2010 (công trình số [1]) và năm 2013 (công trình số [4]); kỷ yếu hội thảo quốc gia lần thứ XVI *"Một số vấn đề chọn lọc của Công nghệ thông tin và Truyền thông"* năm 2013 (công trình số [6]); tạp chí khoa học Trường Đại học Quy Nhơn năm 2014 (công trình số [7]).

- Đề xuất một số cải tiến đối với mô hình gióng hàng IBM theo cách tiếp cận dựa trên ràng buộc, bao gồm: ràng buộc neo, ràng buộc về vị trí của từ, ràng buộc về từ loại và ràng buộc về cụm từ. Với mỗi ràng buộc, chúng tôi đưa ra phương pháp tổng quát để tích hợp nó vào thuật toán EM trong quá trình ước lượng tham số của mô hình. Ngoài ra, chúng tôi đưa ra một phương pháp để kết hợp các ràng buộc. Những cải tiến này đã giúp nâng cao chất lượng dịch cho hệ thống dịch máy thống kê Anh - Việt. Đóng góp này đã được công bố ở kỷ yếu hội thảo quốc tế *International Conference on Asian Language Processing (IALP)* năm 2012 (công trình số [2]); kỷ yếu hội thảo quốc gia lần thứ XV *"Một số vấn đề chọn lọc của Công nghệ thông tin và Truyền thông"* năm 2012 (công trình số [3]); tạp chí *The International Journal of Knowledge and Systems Science (IJKSS)* năm 2014 (công trình số [8]).

- Đề xuất phương pháp xác định cụm từ song ngữ cho dịch máy thống kê. Chúng tôi sử dụng các mẫu cú pháp kết hợp với gióng hàng cụm từ để

xác định cụm từ song ngữ. Các cụm từ song ngữ này đã được ứng dụng vào việc nâng cao chất lượng dịch cho hệ thống dịch máy thống kê Anh - Việt. Đóng góp này đã được công bố ở kỷ yếu hội thảo quốc tế *Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RIVF)* năm 2013 (công trình số [5]).

Các nội dung và kết quả nghiên cứu trình bày trong luận án (từ Chương 2 đến Chương 4) đã được công bố trong 8 công trình. Trong đó, 1 bài báo ở tạp chí quốc tế có phản biện, được xuất bản bởi IGI Global; 4 báo cáo trong kỷ yếu của hội nghị quốc tế có phản biện, được xuất bản bởi IEEE và Springer; 2 báo cáo trong kỷ yếu của hội thảo quốc gia có phản biện và 1 bài báo ở tạp chí trong nước có phản biện.

4. Bố cục của luận án

Ngoài phần mở đầu và kết luận, luận án được tổ chức thành 4 chương, với bố cục như sau:

- **Chương 1.** Giới thiệu tổng quan về các vấn đề nghiên cứu trong luận án. Chúng tôi phân tích, đánh giá các công trình nghiên cứu liên quan; nêu ra một số vấn đề còn tồn tại mà luận án sẽ tập trung giải quyết; xác định nội dung nghiên cứu của luận án.
- **Chương 2.** Trình bày nội dung, kết quả nghiên cứu về xây dựng ngữ liệu song ngữ cho dịch máy thống kê.
- **Chương 3.** Trình bày nội dung, kết quả nghiên cứu về một số cải tiến mô hình IBM để gióng hàng từ cho dịch máy thống kê.
- **Chương 4.** Trình bày nội dung, kết quả nghiên cứu về xác định cụm từ song ngữ cho dịch máy thống kê.

Chương 1

Tổng quan

Chương này trình bày tổng quan về các vấn đề nghiên cứu trong luận án, bao gồm: khai phá tri thức song ngữ, sơ lược về dịch máy (Machine Translation - MT) và dịch máy thống kê (Statistical Machine Translation - SMT). Tiếp đến, chúng tôi phân tích, đánh giá các công trình nghiên cứu liên quan. Cuối chương, chúng tôi nêu ra một số vấn đề còn tồn tại mà luận án sẽ tập trung giải quyết và xác định nội dung nghiên cứu của luận án.

1.1 Khai phá tri thức song ngữ

Nhiệm vụ của khai phá tri thức song ngữ (mining parallel knowledge) là tự động tìm ra các thành phần có ngữ nghĩa tương ứng trong các văn bản ở hai ngôn ngữ khác nhau. Tri thức song ngữ gồm nhiều khía cạnh: song ngữ về từ, song ngữ về cụm từ, song ngữ về cấu trúc, vv. Việc khai phá tri thức song ngữ là quá trình chuẩn bị và khai phá dữ liệu cho một số ứng dụng quan trọng trong lĩnh vực xử lý ngôn ngữ tự nhiên (Natural Language Processing - NLP), trong đó có SMT. Trong luận án này, chúng tôi giới hạn việc khai phá tri thức song ngữ cho bài toán SMT. Sau đây, chúng tôi sẽ trình bày tổng quan về xây dựng ngữ liệu song ngữ, gióng hàng văn bản và xác định cụm từ song ngữ.

Bảng 1.1: Ví dụ về một văn bản song ngữ Anh - Việt.

Văn bản tiếng Anh	**Văn bản tiếng Việt**
In the early summer of 2004, I got a phone call from Steve Jobs. He had been scattershot friendly to me over the years, with occasional bursts of intensity, especially when he was launching a new product that he wanted on the cover of Time or featured on CNN, places where I'd worked.	Đầu mùa hè năm 2004, tôi nhận được một cuộc gọi từ Steve Jobs. Jobs chỉ liên lạc với tôi khi có việc cần trong nhiều năm qua, và có lúc tôi bị ông khủng bố điện thoại, đặc biệt là khi chuẩn bị ra mắt một sản phẩm mới và muốn nó nằm ngay trên trang bìa của tạp chí Time hoặc trình chiếu trên CNN, nơi tôi làm việc.
But now that I was no longer at either of those places, I hadn't heard from him much. We talked a bit about the Aspen Institute, which I had recently joined, and I invited him to speak at our summer campus in Colorado. He'd be happy to come, he said, but not to be onstage. He wanted instead to take a walk so that we could talk.	Nhưng giờ tôi không chẳng còn làm ở cả hai nơi đó nữa và cũng không nghe tin về ông nhiều. Chúng tôi đã trao đổi qua về học viện Aspen, nơi tôi mới vào làm lúc đó, và tôi đã mời ông đến phát biểu tại trại hè của chúng tôi ở Colorado, ông vui vẻ nhận lời đến tham dự nhưng sẽ không lên phát biểu, thay vào đó chúng tôi sẽ nói chuyện trong khi đi dạo.
I had known him since 1984, when he came to Manhattan to have lunch with Time's editors and extol his new Macintosh. He was petulant even then, attacking a Time correspondent for having wounded him with a story that was too revealing.	Tôi quen ông từ năm 1984, khi ông đến Manhattan để ăn trưa cùng với những biên tập viên của tạp chí Time và nhân tiện giới thiệu luôn chiếc máy Macintosh (Mac) mới của mình. Thậm chí lúc đó ông đã nổi nóng, và tấn công một phóng viên của tạp chí Time vì đã làm ông tổn thương bằng một câu chuyện quá lố.

1.1.1 Xây dựng ngữ liệu song ngữ

Ngữ liệu song ngữ (parallel corpus hoặc parallel corpora[1]) là tập hợp các văn bản song ngữ, Bảng 1.1 trình bày ví dụ về một văn bản song ngữ Anh - Việt. Theo Westerhout [89], trường hợp đơn giản nhất ngữ liệu chỉ gồm hai ngôn ngữ, ví dụ: ngữ liệu Compara [34]. Một số ngữ liệu song ngữ gồm nhiều ngôn ngữ, ví dụ: ngữ liệu Europarl [59] bao gồm các phiên bản của 11 ngôn ngữ châu Âu (trong đó một ngôn ngữ là tiếng Anh) như mô tả trong Bảng 1.2.

Ngữ liệu song ngữ tồn tại theo một số định dạng khác nhau. Nó có thể là văn bản song ngữ ở dạng thô hoặc đã được gióng hàng (alignment). Văn bản song ngữ có thể được gióng hàng ở mức đoạn, mức câu, mức cụm từ hoặc mức từ [15]. Việc

[1]Trong tiếng Anh, *corpora* là hình thức số nhiều của *corpus*.

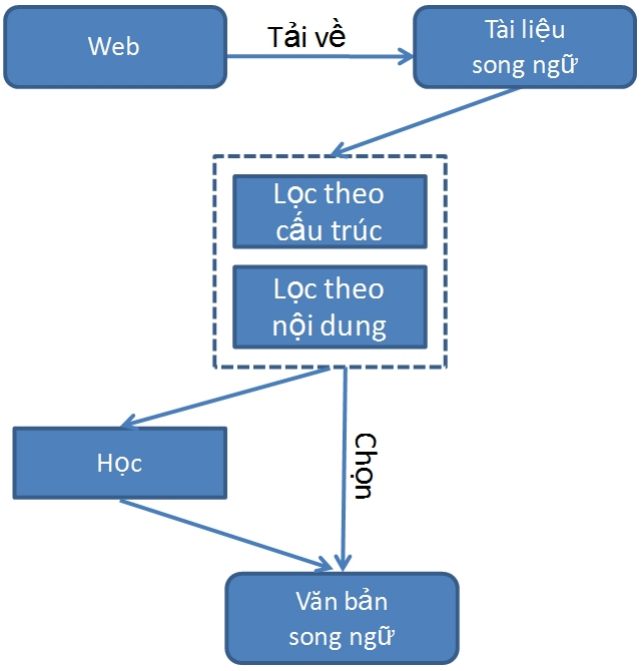
Bảng 1.2: Ngữ liệu Europarl: gồm 10 cặp ngôn ngữ trong đó một ngôn ngữ là tiếng Anh. Ký hiệu L_1 là ngôn ngữ nguồn, L_2 là ngôn ngữ đích.

Ngữ liệu (L_1-L_2)	**Số câu**	**Số từ trong L_1**	**Số từ trong L_2**
Đan Mạch - Anh	1.684.664	43.692.760	46.282.519
Đức - Anh	1.581.107	41.587.670	43.848.958
Hy Lạp - Anh	960.356	-	27.468.389
Tây Ban Nha - Anh	1.689.850	48.860.242	46.843.295
Phần Lan - Anh	1.646.143	32.355.142	45.136.552
Pháp - Anh	1.723.705	51.708.806	47.915.991
Ý - Anh	1.635.140	46.380.851	47.236.441
Hà Lan - Anh	1.715.710	47.477.378	47.166.762
Bồ Đào Nha - Anh	1.681.991	47.621.552	47.000.805
Thụy Điển - Anh	1.570.411	38.537.243	42.810.628

gióng hàng các văn bản song ngữ rất hữu ích cho các ứng dụng khác nhau trong NLP. Các hệ thống SMT [10] sử dụng câu song ngữ làm đầu vào cho mô-đun gióng hàng từ để thực hiện tính toán xác suất dịch từ. Các hệ thống truy vấn thông tin liên ngữ [25, 90, 118] sử dụng văn bản song ngữ để xác định thông tin tương ứng trong cả hai giai đoạn hỏi và đáp. Ngoài ra, việc rút trích các thành phần ngữ nghĩa tương đương của các văn bản song ngữ như từ, cụm từ và câu rất hữu ích cho việc xây dựng từ điển song ngữ [65, 78]. Trong luận án này, chúng tôi giới hạn việc xây dựng ngữ liệu song ngữ cho SMT.

Ngày nay, cùng với sự phát triển của Internet, Web là nguồn cơ sở dữ liệu khổng lồ chứa các tài liệu đa ngôn ngữ (multi-language), nguồn dữ liệu này được sử dụng cho các ứng dụng xử lý văn bản song ngữ. Vì lý do này, nhiều nghiên cứu tập trung vào việc rút trích dữ liệu song ngữ tự động từ Web. Về cơ bản, chúng tôi có thể phân loại các nghiên cứu này vào ba nhóm: (i) dựa trên nội dung (content - based) [16, 24, 76], (ii) dựa trên cấu trúc (structure - based) [17, 97, 100] và (iii) kết hợp (i) với (ii) [101, 128]. Hình 1.1 trình bày sơ đồ tổng quan về rút trích ngữ liệu song ngữ từ Web.

Cách tiếp cận dựa trên nội dung thường dùng từ điển song ngữ để đo độ tương tự về nội dung của hai văn bản. Khi từ điển song ngữ có sẵn, tài liệu ở ngôn ngữ nguồn được dịch theo từng từ (word by word) ra ngôn ngữ đích. Các tài liệu dịch này sau đó được sử dụng để tìm tài liệu song ngữ phù hợp nhất bằng cách sử dụng các đo độ như Cosine, Jaccard, Dice, vv [55]. Tuy nhiên, sử dụng từ điển song ngữ có thể phải đối mặt với khó khăn vì một từ thường có nhiều bản dịch của nó. Để



Hình 1.1: Sơ đồ tổng quan về rút trích ngữ liệu song ngữ từ Web.

khắc phục hạn chế này, chúng tôi sử dụng một hệ thống SMT để có thể tận dụng những lợi thế của phương pháp dịch thống kê trong việc giải quyết các vấn đề về nhập nhằng từ vựng.

Cách tiếp cận dựa trên cấu trúc so khớp cấu trúc HTML (HyperText Markup Language) của trang *web*. Cách tiếp cận này sử dụng giả thuyết các trang *web* song ngữ được trình bày với cấu trúc tương tự nhau. Hệ thống STRAND của Resnik [101] là đại diện tiêu biểu cho cách tiếp cận này. Độ tương tự của các trang *web* được tính dựa vào cấu trúc HTML của chúng. Lưu ý rằng, các phương pháp dựa trên cách tiếp cận này không đòi hỏi tri thức về ngôn ngữ và khá hiệu quả trong việc loại ra các cặp tài liệu không phải song ngữ. Tuy nhiên, nó có hạn chế là yêu cầu hai trang *web* song ngữ phải có cùng một cách trình bày. Theo quan sát của chúng tôi, nhiều trang *web* sử dụng cùng một mẫu thiết kế *web*, vì thế cấu trúc của các trang tương tự nhưng nội dung của chúng lại khác nhau. Do đó, phương pháp tiếp cận dựa trên cấu trúc HTML không được áp dụng trong một số trường hợp. chúng tôi đã kết hợp các đặc trưng dựa trên nội dung với các đặc trưng dựa trên cấu trúc của trang *web* để rút trích các văn bản song ngữ. Để tăng độ chính xác trong việc rút trích các văn bản song ngữ từ Web, chúng tôi kết hợp cả đặc trưng về cấu trúc và đặc trưng về nội dung[2].

[2]Chi tiết chúng tôi trình bày trong Chương 2, phần 2.1

Hiện tại, có ít nghiên cứu về vấn đề này liên quan đến cặp ngôn ngữ Anh - Việt. Hai tác giả Đặng Bác Văn và Hồ Bảo Quốc [24] xây dựng ngữ liệu song ngữ Anh - Việt dựa trên việc so khớp nội dung. Trước hết, các cặp trang *web* ứng viên được xác định bằng cách sử dụng các đặc trưng về độ dài câu và ngày tạo trang *web*. Sau đó, các tác giả đo độ tương tự về nội dung sử dụng từ điển song ngữ Anh - Việt để quyết định hai trang *web* có phải là song ngữ hay không. Quá trình này được thực hiện dựa trên một số ngưỡng của độ đo này. Chú ý rằng, phương pháp trong [24] chỉ tìm kiếm các trang *web* song ngữ có chất lượng dịch tốt và các trang song ngữ này có cùng kiểu trình bày. Hơn nữa, sử dụng từ điển để dịch theo từng từ có thể gây ra sự nhập nhằng. Vì thế, cách tiếp cận này khó để mở rộng khi dữ liệu tăng lên hoặc các trang song ngữ có kiểu trình bày khác nhau.

Như chúng tôi đã đề cập ở trên, Web là nguồn cơ sở dữ liệu khổng lồ chứa các tài liệu đa ngôn ngữ. Tuy nhiên, để có được ngữ liệu song ngữ với độ chính xác cao vẫn đang là một thách thức, bởi vì các văn bản được trình bày trên Internet thường bị "nhiễu". Trong khi đó, nhiều sách điện tử song ngữ (sẵn có) chứa một số lượng lớn các văn bản song ngữ được dịch cẩn thận. Đây là nguồn dữ liệu rất tiềm năng để bổ sung ngữ liệu song ngữ cho SMT, đặc biệt đối với các cặp ngôn ngữ còn hạn chế về ngữ liệu song ngữ như Anh - Việt, Nhật - Việt, vv. Hiện tại, các ngữ liệu song ngữ có sẵn không những có kích thước tương đối nhỏ mà còn không cân bằng ngay cả đối với các ngôn ngữ chính [24], điều này ảnh hưởng đến chất lượng của các hệ thống SMT.

1.1.2 Gióng hàng văn bản

Trong xử lý văn bản song ngữ, gióng hàng là bài toán quan trọng nhất, tức là phát hiện sự tương ứng giữa các đơn vị trong hai văn bản ở các ngôn ngữ khác nhau [4]. Gióng hàng có thể được thực hiện ở mức đoạn, câu, cụm từ hoặc từ. Trong luận án này, chúng tôi giới hạn ở ba mức gióng hàng, cụ thể là: gióng hàng đoạn và gióng hàng câu để xây dựng ngữ liệu và gióng hàng từ cho SMT.

1.1.2.1 Gióng hàng đoạn/câu

Về cơ bản, gióng hàng đoạn và gióng hàng câu có cách tiếp cận tương tự nhau. Để tăng độ chính xác, chúng ta có thể gióng hàng đoạn trước rồi sau đó gióng hàng câu. Việc gióng hàng đoạn đặc biệt quan trọng khi các văn bản cần gióng hàng có

kích thước lớn, ví dụ như sách điện tử. Nhiệm vụ của gióng hàng đoạn/câu là liên kết các đoạn/câu trong một văn bản ở ngôn ngữ này (ngôn ngữ nguồn) với các đoạn/câu là bản dịch tương ứng của nó trong một văn bản ở ngôn ngữ khác (ngôn ngữ đích) [21]. Các phương pháp khác nhau đã được đề xuất cho việc xác định gióng hàng đoạn/câu giữa các văn bản song ngữ [41, 98, 114]. Theo quan điểm của chúng tôi, những phương pháp này có thể được chia thành hai cách tiếp cận chính: (i) dựa trên thống kê (statistics - based) [11, 35] và (ii) dựa trên tri thức ngôn ngữ (linguistic knowledge - based) [18, 80].

Cách tiếp cận thứ nhất (i) khai thác các mối tương quan về độ dài của các khối văn bản (đoạn hoặc câu) trong các ngôn ngữ khác nhau và cố gắng thiết lập sự tương ứng giữa các khối văn bản này theo kích thước [37]. Ở đây, kích thước có thể được đo bởi số từ hoặc số ký tự. Gale và cộng sự [35] đã sử dụng mô hình thống kê đơn giản theo độ dài với kích thước là số từ để gióng hàng câu cho ngữ liệu song ngữ. Trong mô hình này, mỗi cặp câu được gán một xác suất. Xác suất này được sử dụng để tìm khả năng liên kết cực đại của các câu (dựa trên kỹ thuật quy hoạch động). Tuy nhiên, các văn bản thường được định dạng lại trong quá trình dịch thuật. Vì vậy, nó không chỉ chứa các liên kết *1-1*, tức là một đoạn/câu trong văn bản ở ngôn ngữ nguồn có thể liên kết với hai hoặc nhiều đoạn/câu trong văn bản ở ngôn ngữ đích và ngược lại. Trong trường hợp này, phương pháp thống kê dựa vào cấu trúc như từ hoặc ký tự có thể không thực hiện tốt.

Cách tiếp cận thứ hai (ii) sử dụng dữ liệu ngôn ngữ (thường là từ điển) để thiết lập sự tương ứng giữa các khối văn bản. Li và cộng sự [68] đề xuất thuật toán Fast-Champollion, trong đó sử dụng từ điển song ngữ cho việc gióng hàng câu. Với thuật toán này, độ chính xác (precision) và độ bao phủ (recall) phụ thuộc vào kích thước của từ điển được sử dụng. Ngoài ra, làm thế nào để xây dựng từ điển song ngữ tự động là một vấn đề quan trọng đối với việc áp dụng thuật toán Fast-Champollion trên các cặp ngôn ngữ không có sẵn từ điển[3].

1.1.2.2 Gióng hàng từ

Gióng hàng từ (word alignment) là một nhiệm vụ xác định sự tương ứng giữa các từ trong một văn bản song ngữ [72]. Đây là bước đầu tiên trong hầu hết các cách tiếp cận hiện tại của SMT. Ayan [4] đã chỉ ra rằng, chất lượng của gióng hàng từ đóng vai trò rất quan trọng cho sự thành công của một hệ thống SMT. Các

[3]Ở đây, chúng tôi muốn nói đến từ điển song ngữ điện tử.

phương pháp khác nhau đã được đề xuất để xác định gióng hàng từ trong các văn bản song ngữ. Nói chung, các phương pháp gióng hàng từ có thể được phân chia thành hai loại: (i) cách tiếp cận dựa trên mô hình phân biệt (discriminative model) và (ii) cách tiếp cận dựa trên mô hình sinh (generative model).

Cách tiếp cận thứ nhất (i) dựa vào quá trình huấn luyện trên một tập các đặc trưng, điển hình là các nghiên cứu của Moore [83] và Liu [72]. Cách tiếp cận này có ưu điểm là linh hoạt trong việc kết hợp các đặc trưng mới [77]. Tuy nhiên, hạn chế của cách tiếp cận này là dữ liệu huấn luyện cần phải được gán nhãn; công việc này đòi hỏi nhiều thời gian, chi phí để thực hiện và nó không sẵn có với hầu hết các cặp ngôn ngữ [74]. Ngoài ra, rất khó khăn để chọn dữ liệu đại diện cho việc huấn luyện để đảm bảo rằng các mô hình sẽ hoạt động tốt trên dữ liệu không quan sát được, đặc biệt khi dữ liệu song ngữ đến từ nhiều nguồn thuộc nhiều lĩnh vực khác nhau [72].

Cách tiếp cận thứ hai (ii) thường sử dụng mô hình sinh, trong đó các mô hình IBM của Brown và cộng sự [12] được sử dụng rộng rãi nhất. Thuật toán cực đại kỳ vọng (Expectation Maximization - EM) [27] được sử dụng để ước lượng xác suất của mô hình gióng hàng trên ngữ liệu song ngữ. Các mô hình này về cơ bản là độc lập với ngôn ngữ và các tham số của nó được ước lượng từ ngữ liệu với tối thiểu việc tiền xử lý [111]. Tuy nhiên, chất lượng của gióng hàng thường khá thấp đối với các cặp ngôn ngữ có nhiều khác biệt về cấu trúc cú pháp như Anh - Việt, Anh - Trung, vv. Vì vậy, sử dụng thêm các nguồn tri thức bên ngoài như thông tin về từ vựng, thông tin về cú pháp là thật sự cần thiết để cải thiện chất lượng của gióng hàng.

Trong các nghiên cứu trước đây, các mô hình IBM được cải tiến với nhiều phương pháp khác nhau. Varea và cộng sự [115] sử dụng mô hình *Maximum Entropy* (ME) phụ thuộc ngữ cảnh để chứa nhiều hơn các phụ thuộc. Tức là, một ngữ cảnh lớn hơn được sử dụng trong mô hình dịch thay vì chỉ sử dụng xác suất dịch từ. Một cải tiến khác đối với các mô hình IBM dựa trên mô hình từ vựng đối xứng được đề xuất bởi Zens và cộng sự [125]. Họ áp dụng phương pháp nội suy tuyến tính (linear interpolation) để tính xác suất theo hai hướng (hướng dịch chuẩn từ ngôn ngữ nguồn sang ngôn ngữ đích và hướng dịch ngược lại). Ngoài ra, các tác giả đã mô tả quá trình làm trơn (smoothing) từ vựng bằng cách sử dụng hình thức từ gốc (word base form). Đặc biệt cho các ngôn ngữ biến cách cao (inflected language) như tiếng Đức, điều này dẫn đến những cải tiến đáng kể về mặt thống kê. Moore [82] đã khảo sát ba phương pháp đơn giản để cải tiến mô

hình IBM 1: (i) gắn trọng số cho xác suất gióng hàng với từ rỗng (hay còn gọi là từ *null*), (ii) làm trơn quá trình ước lượng xác suất cho các từ hiếm và (iii) sử dụng phương pháp ước lượng dựa trên kinh nghiệm (heuristic) để khởi tạo hoặc thay thế trong quá trình huấn luyện các tham số của mô hình. Các kết quả thực nghiệm của tác giả với ngữ liệu Anh - Pháp cho thấy tỷ lệ lỗi gióng hàng giảm khi áp dụng ba phương pháp này. Như vậy, trong các nghiên cứu liên quan về cải tiến các mô hình IBM như chúng tôi đã trình bày, mỗi nghiên cứu đưa ra một (hoặc một số) phương pháp khác nhau. Tuy nhiên, trong các nghiên cứu này, các tác giả chưa sử dụng nguồn tri thức mở rộng (ngoài ngữ liệu song ngữ dùng để huấn luyện) vào quá trình gióng hàng.

Nhiều nghiên cứu tập trung vào việc sử dụng các thông tin về từ loại để nâng cao độ chính xác của gióng hàng. Một số thực hiện ở giai đoạn tiền xử lý [38, 124] hoặc hậu xử lý [20, 67] dữ liệu cho các mô hình thống kê. Koehn cùng cộng sự [58] đề xuất mô hình dịch bổ sung tham số ngôn ngữ học (factored translation model), mô hình này cho phép người dùng thêm các lớp thông tin về ngôn ngữ (ví dụ như hình thái từ, nhãn từ loại, vv) vào hệ thống SMT dựa trên cụm từ. Trong mô hình này, dữ liệu huấn luyện được chú thích với các yếu tố bổ sung. Các tác giả đã chỉ ra hiệu suất của SMT đã được cải thiện bằng cách sử dụng các yếu tố này. Tuy nhiên, việc bổ sung các yếu tố ngôn ngữ trực tiếp vào dữ liệu huấn luyện sẽ làm tăng thêm số từ vựng, do đó có thể làm cho dữ liệu huấn luyện thưa hơn.

Đối với cách tiếp cận ràng buộc, một số nghiên cứu đã đề xuất các phương pháp khác nhau để nâng cao chất lượng gióng hàng từ. Lin và Cherry [69] trình bày ràng buộc dựa trên cú pháp để gióng hàng từ, được gọi là ràng buộc "dính liền" (cohesion constraint). Ràng buộc này đòi hỏi các cụm từ tiếng Anh rời nhau được ánh xạ tới các khoảng không giao nhau (non-overlapping) trong câu tiếng Pháp. Nghiên cứu của Kamigaito [52] sử dụng ràng buộc về tần suất (frequency constraint) cho các từ chức năng (function word) và từ nội dung (content word). Với việc sử dụng ràng buộc này, xác suất dịch của mỗi cặp từ được điều chỉnh thông qua tham số λ ở trong thuật toán EM. Các thực nghiệm được tiến hành trên hệ thống SMT Nhật - Anh cho thấy chất lượng MT tăng trung bình 0,2 điểm BLEU khi so sánh với mô hình gốc.

Gần đây, Songyot và cộng sự trong [110] đã chỉ ra một hạn chế của các mô hình IBM, đó là các gióng hàng lỗi xuất xảy ra với các từ có tần số xuất hiện thấp trong dữ liệu huấn luyện. Vấn đề này có thể tồi tệ hơn đối với các ngôn ngữ có ít ngữ liệu song ngữ. Các kỹ thuật làm trơn như của Zhang và Chiang [126] hoặc các

phân bố tiên nghiệm (prior distribution) đã được Vaswani [116] và Mermer [79] sử dụng để giải quyết hạn chế này. Nghiên cứu của Songyot và cộng sự trong [110] sử dụng thông tin học mô hình tương tự từ (word similarity model) từ dữ liệu đơn ngữ dựa trên mạng nơ-ron. Thông tin này sau đó được tích hợp vào các mô hình IBM, kết quả thực nghiệm cho thấy cải thiện đáng kể chất lượng gióng hàng và chất lượng MT trên hai cặp ngôn ngữ Trung - Anh và Ả-rập - Anh. Ngoài ra, một số mô hình gióng hàng không giám sát (unsupervised) giống như các mô hình IBM được đề xuất bởi một số tác giả như Dyer [33], Yang [122], Tamura [112], tuy nhiên nó không được sử dụng rộng rãi như các mô hình IBM.

Một hướng nghiên cứu khác tập trung vào gióng hàng từ dựa trên mô hình phân biệt. Các mô hình lô-ga-rít tuyến tính (log-linear) được đề xuất bởi Liu và cộng sự [70] cho phép mô hình thống kê có thể được mở rộng bằng cách tích hợp thêm các phụ thuộc cú pháp. Ittycheriah [50] trình bày thuật toán gióng hàng từ cho cặp ngôn ngữ Ả-rập - Anh dựa trên mô hình ME sử dụng dữ liệu huấn luyện có gán nhãn. Phương pháp học mô hình gióng hàng từ trên cơ sở các đặc trưng tùy ý của các cặp từ được Taskar trình bày trong [113]. Một số nghiên cứu kết hợp giữa hai cách tiếp cận (mô hình phân biệt và mô hình sinh) như Berg và cộng sự [8], Dyer [32] cho thấy kết quả khả quan.

Việc kết hợp các nguồn tri thức bên ngoài vào quá trình gióng hàng đã được một số tác giả quan tâm nghiên cứu. Och và Ney [92] sử dụng từ điển song ngữ như là nguồn bổ sung tri thức cho việc mở rộng ngữ liệu huấn luyện. Họ gán các cặp từ trong điển đồng thời xuất hiện trong ngữ liệu huấn luyện với trọng số cao và các cặp từ còn lại được gán với trọng số rất thấp. Talbot [111] đề xuất phương pháp sử dụng các nguồn thông tin phụ trợ như các quan hệ *cognate*, từ điển song ngữ, các mẫu so khớp cho các chữ số để hạn chế các gióng hàng không mong muốn. Trong các nghiên cứu này, chưa có phương pháp tổng quát để thêm nguồn tri thức mới và kết hợp chúng lại với nhau.

1.1.3 Xác định cụm từ song ngữ

Các cụm từ song ngữ hữu ích cho nhiều nhiệm vụ trong NLP như truy xuất thông tin liên ngữ [1], phân tích cú pháp [3], khai phá văn bản [102] và đặc biệt là cho SMT [99]. Trong các hệ thống SMT, chất lượng của các bản dịch phụ thuộc chủ yếu vào chất lượng của các cặp cụm từ song ngữ được rút trích từ ngữ liệu song ngữ [117]. Vì vậy, nhiều phương pháp đã được đề xuất để rút trích các cụm từ song

ngữ từ ngữ liệu song ngữ hoặc ngữ liệu có thể so sánh được (comparable corpora) [5, 28]. Theo quan điểm của chúng tôi, những phương pháp này có thể được phân loại thành ba cách tiếp cận chính: "tượng trưng" (symbolic), thống kê (statistics) và phương pháp lai (hybrid).

Cách tiếp cận đầu tiên sử dụng một bộ lọc ngôn ngữ, nó phụ thuộc vào các mẫu cú pháp (syntactic pattern) [96]. Tuy nhiên, rất khó để áp dụng phương pháp "tượng trưng" cho dữ liệu không có chú thích về cú pháp [2, 28]. Cách tiếp cận thứ hai sử dụng các độ đo thống kê như thông tin tương hỗ (mutual information) [127], tỷ lệ lô-ga-rít thích hợp (log-likelihood ratio) [23] để xếp hạng các ứng viên cho cụm từ song ngữ. Ưu điểm chính của phương pháp thống kê là độc lập ngôn ngữ. Tuy nhiên, hạn chế của cách tiếp cận này là phải có được một ngữ liệu đủ lớn. Ngoài ra, các độ đo thống kê chủ yếu được áp dụng cho *2-gram* và *3-gram* và nó sẽ trở nên khó khăn hơn khi rút trích các cụm từ nhiều hơn ba từ [2]. Cách tiếp cận thứ ba kết hợp cả hai cách tiếp cận trước [108]. Cách tiếp cận này rút trích các ứng viên của cụm từ song ngữ sử dụng một bộ lọc ngôn ngữ, sau đó gán cho mỗi ứng viên của cụm từ song ngữ một điểm số tùy thuộc vào phương pháp thống kê [54].

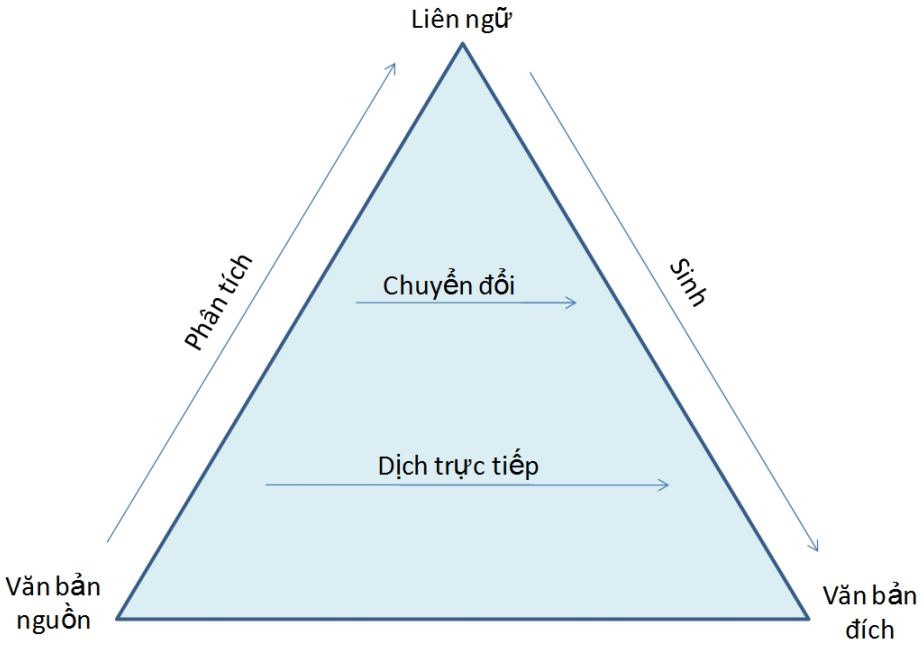
Trong các nghiên cứu liên quan sử dụng mẫu cú pháp để xác định cụm từ song ngữ. Việc so khớp các mẫu cú pháp được thực hiện ở hai phía (cả câu nguồn và câu đích). Với cách làm này, chúng ta chỉ rút trích được các cụm từ song ngữ với số lượng hạn chế. Bouamor và cộng sự [9] đã chỉ ra rằng, các cụm từ song ngữ được sử dụng để cải thiện chất lượng dịch cho SMT.

1.2 Sơ lược về dịch máy

Không lâu sau khi những chiếc máy tính điện tử đầu tiên[4] ra đời, Warren Weaver[5] (1949) đưa ra ý tưởng rằng, có thể một ngày nào đó máy tính nhận đầu vào là một tài liệu viết bằng một số ngôn ngữ nào đó (ngôn ngữ nguồn) và tự động tạo ra một tài liệu tương đương viết bằng một số ngôn ngữ khác (ngôn ngữ đích) - một nhiệm vụ mà bây giờ chúng ta gọi là MT. Từ đó đến nay, sau hơn 60 năm nghiên cứu và phát triển, các dịch vụ MT bây giờ đã trở nên phổ biến rộng rãi và được sử dụng miễn phí, nó nhận được hàng trăm triệu yêu cầu mỗi tuần [42].

[4]ENIAC - Máy tính điện tử đầu tiên ra đời năm 1946 [75].

[5]Tiến sĩ Warren Weaver (17/7/1894 - 24/11/1978), là một nhà khoa học người Mỹ. Ông là một trong những người đi tiên phong về MT [60].



Hình 1.2: Kim tự tháp dịch máy.

Tính hữu ích của công nghệ MT tăng lên cùng với chất lượng dịch. Theo Koehn [60], việc sử dụng MT có thể được chia thành ba loại: (i) dịch tài liệu tiếng nước ngoài cho mục đích hiểu nội dung, (ii) dịch văn bản để xuất bản ở các ngôn ngữ khác và (iii) thông tin liên lạc, chẳng hạn như dịch *email*, *chat*, vv. Mỗi một ứng dụng đòi hỏi tốc độ và chất lượng khác nhau.

Một số tiêu chí có thể được sử dụng để phân loại các cách tiếp cận MT, nhưng tiêu chí phân loại phổ biến nhất được sử dụng là mức độ phân tích ngôn ngữ (linguistic analysis) theo yêu cầu của hệ thống để tạo ra các bản dịch. Thông thường, điều này có thể được thể hiện một cách trực quan bằng sơ đồ "kim tự tháp dịch máy" (machine translation pyramid) như mô tả trong Hình 1.2.

Trước kỹ thuật dịch thống kê, có bốn cách tiếp cận cho bài toán MT [51], bao gồm: dịch trực tiếp [53], dịch dựa trên chuyển đổi [66], dịch liên ngữ [81] và dịch dựa trên ví dụ [87, 103]. Trong cách dịch trực tiếp, quá trình dịch được thực hiện từng từ một bằng cách sử dụng từ điển song ngữ lớn và sắp xếp lại thứ tự các từ theo các quy tắc cho trước. Cách tiếp cận chuyển đổi dựa vào việc phân tích một câu trước khi dịch, sau đó dịch cấu trúc câu và tạo ra một câu trong ngôn ngữ khác. Cách tiếp cận thứ ba là phân tích các thông tin của câu để tạo thành một biểu diễn ý nghĩa trừu tượng, điều này được biết đến như là một "ngôn ngữ quốc tế" (hay liên ngữ - interlingua) trước khi tạo ra một câu trong ngôn ngữ khác. Đối

với cách tiếp cận dựa trên ví dụ, hệ thống dịch tìm câu tương tự với câu đầu vào trong ngữ liệu song ngữ (các ví dụ) và thực hiện một số thay đổi thích hợp trong quá trình dịch [60].

1.3 Dịch máy thống kê

Vào cuối những năm 1980, ý tưởng về SMT được ra đời ở phòng thí nghiệm của IBM Research[6] trong bối cảnh thành công của các phương pháp thống kê trong nhận dạng giọng nói [60]. Bằng cách mô hình hóa nhiệm vụ dịch là một bài toán tối ưu hóa thống kê (statistical optimization), dự án Candide [26] đã đặt MT trên một nền tảng toán học vững chắc.

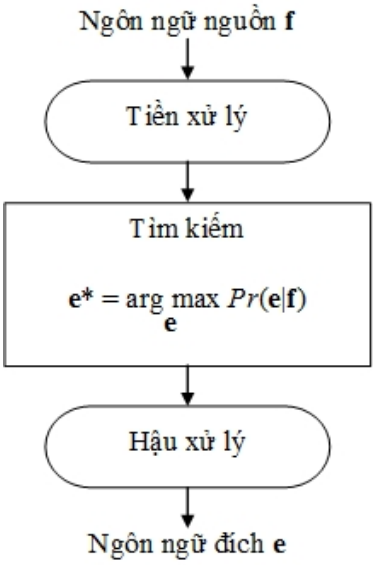
Các hệ thống SMT hiện đang được phát triển mạnh mẽ với một số lượng lớn các phòng thí nghiệm nghiên cứu học thuật. Ngoài ra, nhiều hệ thống SMT thương mại cũng đang được phát triển bởi các công ty phần mềm lớn như IBM, Microsoft và Google. Theo Koehn [60], người sử dụng Internet dịch 50 triệu trang *web* mỗi ngày, sử dụng các hệ thống được cung cấp bởi Google, Yahoo, Microsoft và một số công ty khác.

SMT là một phương pháp MT mà bản dịch được tạo ra trên cơ sở các mô hình thống kê, trong đó các tham số của mô hình được ước lượng từ việc phân tích các ngữ liệu (văn bản đơn ngữ hoặc song ngữ). Thay vì xây dựng các từ điển, các quy luật chuyển đổi bằng tay, hệ dịch này tự động xây dựng các từ điển, các quy luật dựa trên kết quả thống kê có được từ ngữ liệu. Dịch máy dựa trên phương pháp thống kê tìm câu $\mathbf{e}$ ở ngôn ngữ đích phù hợp nhất (có xác suất cao nhất) khi cho trước câu $\mathbf{f}$ ở ngôn ngữ nguồn, như biểu diễn ở công thức (1.1). Hình 1.3 mô hình hoá bài toán MT dựa trên phương pháp thống kê.

$$\mathbf{e}^* = \arg\max_{\mathbf{e}} Pr(\mathbf{e}|\mathbf{f}) \tag{1.1}$$

Phương pháp thống kê có một số ưu điểm so với các phương pháp khác. Các mối quan hệ giữa các từ, cụm từ và cấu trúc ngữ pháp thường không rõ ràng. Các phân bố xác suất và kỹ thuật thống kê cho phép chúng ta xác định điều này [13]. Một mô hình thống kê có thể được huấn luyện trên số lượng lớn dữ liệu và tăng số

[6] http://www.research.ibm.com/



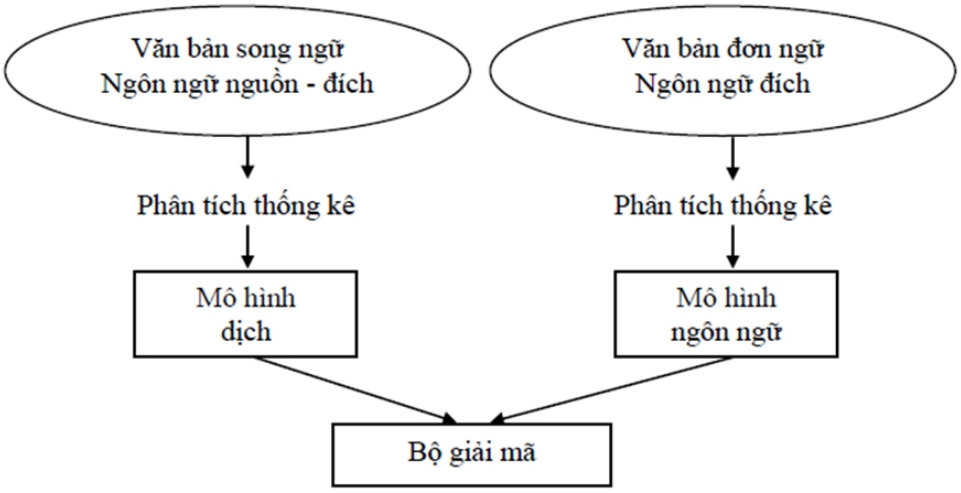
Hình 1.3: Mô hình hoá bài toán dịch máy dựa trên phương pháp thống kê.

lượng dữ liệu huấn luyện sẽ cho phép các mô hình xác định thêm các "hiện tượng ngôn ngữ" (linguistic phenomena) trong các ngôn ngữ. Vì vậy, khi tăng số lượng dữ liệu huấn luyện sẽ dẫn đến các bản dịch chất lượng cao hơn.

Một lợi ích nữa của kỹ thuật thống kê là không cần phải dựa vào các đặc trưng riêng biệt của các ngôn ngữ có liên quan, chẳng hạn như các mô hình ngôn ngữ cụ thể của bản dịch hay ngữ pháp [13]. Nhiều đặc trưng của các mô hình dịch là độc lập ngôn ngữ (language-independent) và có thể được điều chỉnh cho cặp ngôn ngữ cụ thể bằng cách ước lượng các tham số mô hình. Điều này cho phép các hệ thống SMT được xây dựng cho nhiều cặp ngôn ngữ với sửa đổi tối thiểu về mặt kỹ thuật. Để tăng chất lượng dịch, tri thức cụ thể của ngôn ngữ có liên quan thường được yêu cầu. Mô hình thống kê đã được phát triển để kết hợp thông tin ngôn ngữ cụ thể bổ sung tương đối dễ dàng, bao gồm các đặc điểm hình thái, trật tự từ và các mô hình ngữ pháp.

1.3.1 Mô hình hóa bài toán

Nhiệm vụ của một hệ thống SMT là xây dựng mô hình xác suất dịch $Pr(\mathbf{e}|\mathbf{f})$, trong đó câu nguồn $\mathbf{f}$ được dịch sang câu đích $\mathbf{e}$. Brown và cộng sự [12] sử dụng quy tắc Bayes để xây dựng công thức tính xác suất dịch câu nguồn $\mathbf{f}$ sang câu



HÌNH 1.4: Các thành phần của dịch máy thống kê.

đích **e** như sau:

$$\begin{aligned} \mathbf{e}^* &= \arg\max_{\mathbf{e}} Pr(\mathbf{e}|\mathbf{f}) \\ &= \arg\max_{\mathbf{e}} \frac{Pr(\mathbf{f}|\mathbf{e})Pr(\mathbf{e})}{Pr(\mathbf{f})} \\ &= \arg\max_{\mathbf{e}} Pr(\mathbf{f}|\mathbf{e})Pr(\mathbf{e}) \end{aligned} \tag{1.2}$$

Trong đó, $Pr(\mathbf{e})$ là mô hình ngôn ngữ và $Pr(\mathbf{f}|\mathbf{e})$ là mô hình dịch. Mô hình ngôn ngữ $Pr(\mathbf{e})$ được ước lượng từ ngữ liệu ở ngôn ngữ đích (ngữ liệu đơn ngữ) và mô hình dịch $Pr(\mathbf{f}|\mathbf{e})$ được ướng lượng từ ngữ liệu song ngữ. Hình 1.4 trình bày các thành phần của SMT, đó là mô hình dịch, mô hình ngôn ngữ và bộ giải mã.

Mục tiêu của chúng ta ở đây là tìm câu **e** ở ngôn ngữ đích sao cho tích $Pr(\mathbf{f}|\mathbf{e})Pr(\mathbf{e})$ trong công thức (1.2) đạt giá trị cực đại. Ở đây, mô hình dịch $Pr(\mathbf{f}|\mathbf{e})$ được định nghĩa như là xác suất biên (marginal probability), xác suất này bằng tổng tất cả các xác suất gióng hàng từ **a** giữa câu nguồn và câu đích như trong công thức (1.3).

$$Pr(\mathbf{f}|\mathbf{e}) = \sum_{\mathbf{a}} Pr(\mathbf{f}, \mathbf{a}|\mathbf{e}) \tag{1.3}$$

1.3.2 Mô hình ngôn ngữ

Một thành phần quan trọng đối với bất kỳ hệ thống SMT nào, đó là mô hình ngôn ngữ (language model). Rõ ràng, chúng ta muốn có một hệ thống MT không chỉ để

tạo ra các từ đúng về ý nghĩa mà còn để xâu chuỗi chúng lại với nhau thành một câu "trôi chảy" (fluent) ở ngôn ngữ đích. Mô hình ngôn ngữ sẽ hỗ trợ các quyết định khó khăn về trật tự từ (word order) và dịch từ (word translation) [60]. Ví dụ, mô hình ngôn ngữ gán xác suất cao hơn cho câu có trật tự từ đúng so với câu có trật tự từ không đúng: *Pr(ngôi nhà nhỏ) > Pr(nhỏ ngôi nhà).*

Một cách hình thức, mô hình ngôn ngữ là một hàm nhận tham số đầu vào là một câu và trả về xác suất của câu thuộc ngôn ngữ. Ví dụ, ở trong tiếng Việt: *Pr(ngôi nhà nhỏ)* = 0,25; *Pr(nhỏ ngôi nhà)* = 0,01. Như vậy, một mô hình ngôn ngữ tốt sẽ gán xác suất cao hơn cho câu đầu tiên (câu *ngôi nhà nhỏ*). Ưu điểm này của mô hình ngôn ngữ giúp hệ thống SMT xác định được trật tự từ đúng.

Một khía cạnh khác mà mô hình ngôn ngữ mang lại là sự lựa chọn từ. Nếu một từ ở ngôn ngữ nguồn có nhiều bản dịch ở ngôn ngữ đích (chẳng hạn như từ *Haus* trong tiếng Đức dịch sang tiếng Anh là *house*, *home*, ...), xác suất dịch từ vựng sẽ ưu tiên cho bản dịch phổ biến hơn (từ *house*) [60]. Tuy nhiên, trong ngữ cảnh cụ thể, các bản dịch khác có thể được lựa chọn. Tức là, nó cung cấp xác suất cao hơn để lựa chọn từ tự nhiên hơn trong ngữ cảnh cụ thể, ví dụ:

$$Pr(I\ am\ going\ home) > Pr(I\ am\ going\ house) \tag{1.4}$$

Phương pháp hàng đầu cho các mô hình ngôn ngữ là mô hình ngôn ngữ *n-gram*. Mô hình ngôn ngữ *n-gram* dựa trên các số liệu thống kê những từ có khả năng theo sau các từ khác. Ở ví dụ trong công thức (1.4), nếu chúng ta phân tích với một số lượng lớn các văn bản, chúng ta sẽ quan sát thấy từ *home* theo sau từ *going* thường xuyên hơn so với từ *house*.

Trong mô hình ngôn ngữ *n-gram*, chúng ta muốn tính xác suất của câu $s = w_1, w_2, ..., w_n$. Xác suất của câu s được phân rã thành tích của các xác suất có điều kiện. Bằng cách sử dụng quy tắc dây chuyền (chain rule), điều này có thể được thực hiện như trong công thức (1.5). Xác suất của câu $Pr(s)$ được phân rã ra như là xác suất của từng từ riêng lẻ $Pr(w)$.

$$Pr(w_1, w_2, ..., w_n) = Pr(w_1)Pr(w_2|w_1)...Pr(w_n|w_1, w_2, ..., w_{n-1}) \tag{1.5}$$

Để có thể ước lượng được các phân phối xác suất từ trong công thức (1.5), sử dụng xấp xỉ Markov, ta có xác suất xuất hiện của một từ w_n được coi như chỉ phụ

thuộc vào m từ đứng liền trước nó:

$$Pr(w_n|w_1, w_2, ..., w_{n-1}) \simeq Pr(w_n|w_{n-m}, ..., w_{n-2}, w_{n-1}) \tag{1.6}$$

Thông thường, chúng ta chọn giá trị của m dựa trên lượng dữ liệu huấn luyện chúng ta có. Nhiều dữ liệu huấn luyện cho phép giá trị m lớn hơn. Mô hình ngôn ngữ *trigram* được sử dụng phổ biến nhất. Với mô hình này, chúng ta xem xét hai từ đứng trước (tức là $m = 2$) để dự đoán từ thứ ba. Điều này đòi hỏi việc thu thập số liệu thống kê trên các chuỗi ba từ, vì thế được gọi là *3-gram* (trigram). Ngoài ra, các mô hình ngôn ngữ cũng có thể được ước lượng với *2-gram* (bigram), *1-gram* (unigram), vv.

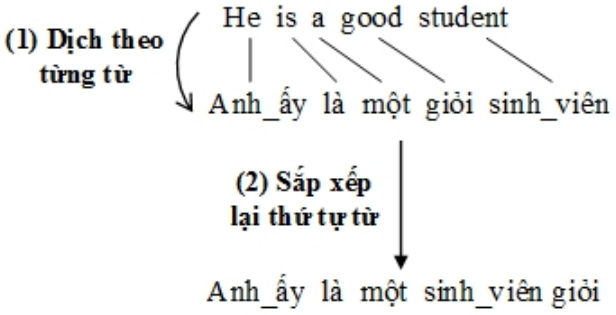
Ở dạng đơn giản nhất, chúng ta ước lượng các xác suất *trigram* là $Pr(w_3|w_1, w_2)$. Để thực hiện công việc này, chúng ta đếm số chuỗi w_1, w_2 được theo sau bởi từ w_3 (ký hiệu $count(w_1, w_2, w_3)$) và số chuỗi w_1, w_2 được theo sau bởi các từ khác (ký hiệu $\sum_w count(w_1, w_2, w)$) trong ngữ liệu huấn luyện. Để ước lượng khả năng cực đại (Maximum Likelihood Estimation - MLE), chúng ta tính:

$$Pr(w_3|w_1, w_2) = \frac{count(w_1, w_2, w_3)}{\sum_w count(w_1, w_2, w)} \tag{1.7}$$

Trong thực tế chúng ta gặp phải vấn đề dữ liệu thưa (data sparseness). Sự phân bố không đều trong tập dữ liệu huấn luyện có thể dẫn đến các ước lượng không chính xác. Khi các *n-gram* phân bố thưa, nhiều cụm *n-gram* không xuất hiện, các cụm *n-gram* này sẽ có xác suất bằng 0. Để khắc phục tình trạng này, người ta sử dụng các kỹ thuật làm trơn (smoothing) nhằm đánh giá chính xác hơn xác suất của các cụm *n-gram*. Một số kỹ thuật làm trơn phổ biến như Add-one, Good – Turing [60], Kneser-Ney [56], vv.

1.3.3 Mô hình dịch

Mô hình dịch (translation model) giúp tính toán xác suất có điều kiện $Pr(\mathbf{f}|\mathbf{e})$. Xác suất này được ước lượng từ ngữ liệu song ngữ của cặp ngôn ngữ nguồn - đích. Có ba hướng tiếp cận chính: (i) dựa trên từ (word - based), (ii) dựa trên cụm từ (phrase - based) và (iii) dựa trên cú pháp (syntax - based).



Hình 1.5: Quá trình dịch dựa trên từ. Câu đầu vào tiếng Anh được dịch từng từ sang tiếng Việt, sau đó sắp xếp lại trật tự từ.

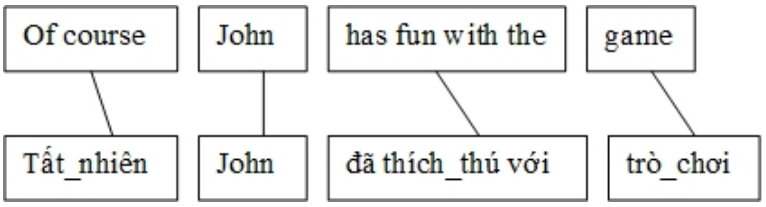
1.3.3.1 Mô hình dịch dựa trên từ

Mô hình dịch dựa trên từ là thế hệ đầu tiên của SMT, được nghiên cứu và phát triển bởi IBM [51]. Với mô hình dịch này, đơn vị được dịch là các từ. Giả sử chúng ta cần dịch câu tiếng Anh *"He is a good student"* sang tiếng Việt theo đơn vị từ. Ví dụ, ta có bản dịch tiếng Việt là *"Anh_ấy là một sinh_viên giỏi"*. Hình 1.5 mô tả ví dụ về quá trình dịch dựa trên từ, gồm 2 bước: (1) dịch theo từng từ: *He* $\rightarrow$ *Anh_ấy*, *is* $\rightarrow$ *là*, *a* $\rightarrow$ *một*, *good* $\rightarrow$ *giỏi*, *student* $\rightarrow$ *sinh_viên*; (2) sắp xếp lại trật tự từ: *Anh_ấy là một giỏi sinh_viên* $\rightarrow$ *Anh_ấy là một sinh_viên giỏi*.

Ở đây, số từ trong câu được dịch là khác nhau phụ thuộc vào các từ ghép, hình thái từ và thành ngữ. Tham số độ dài của chuỗi từ được dịch gọi là độ hỗn loạn (fertility) [57], tức là số từ của ngôn ngữ đích mà từ của ngôn ngữ nguồn sinh ra. Tuy nhiên, tuỳ vào đặc điểm của ngôn ngữ, như cặp ngôn ngữ Anh - Việt cũng giống với cặp ngôn ngữ Anh - Trung, Anh - Nhật, ..., hệ dịch phải đối mặt với khó khăn trong quá trình sắp xếp trật tự của các từ tiếng Anh tương ứng khi dịch sang câu tiếng Việt. Trong quá trình dịch, kết nối từ tiếng Anh tương ứng với từ tiếng Việt có thể là *1-1*, *1-không*, *1-nhiều*, *nhiều-1* hoặc *nhiều-nhiều*. Mô hình dịch dựa trên đơn vị từ không cho kết quả tốt trong trường hợp kết nối *nhiều-1* hoặc *nhiều-nhiều* với trật tự các từ trong câu tương ứng là khác nhau. Khi đó, mô hình dựa trên đơn vị cụm từ được đề xuất để giải quyết vấn đề này.

1.3.3.2 Mô hình dịch dựa trên cụm từ

Cách tiếp cận hiện thành công nhất với SMT là sử dụng cách dịch theo cụm từ. Xem minh họa ở Hình 1.6, trước hết, câu đầu vào tiếng Anh *"Of course John*



Hình 1.6: Dịch dựa trên cụm từ. Câu đầu vào được tách ra thành các cụm từ, dịch một-một các cụm từ tiếng Anh sang tiếng Việt và có thể sắp xếp lại trật tự các cụm từ.

has fun with the game" được tách ra thành các cụm từ: *Of course*, *John*, *has fun with the*, *game*; sau đó, dịch *một-một* các cụm từ tiếng Anh sang tiếng Việt: *Of course* → *Tất_ nhiên*, *John* → *John*, *has fun with the* → *đã thích_ thú với*, *game* → *trò_ chơi*; cuối cùng, có thể sắp xếp lại trật tự các cụm từ này. Ở đây, cụm từ là chuỗi các từ liền kề nhau không nhất thiết là cụm từ trong ngôn ngữ học (theo định nghĩa trong ngữ pháp). Trong phương pháp này, câu đầu vào được chia thành một chuỗi các cụm từ; những cụm từ được ánh xạ *một-một* đến các cụm từ đầu ra, có thể được sắp xếp lại thứ tự các cụm từ. Thông thường, các mô hình cụm từ được ước lượng từ ngữ liệu song ngữ đã được gióng hàng từ. Tất cả các cặp cụm từ nhất quán với gióng hàng từ sẽ được rút trích và gán với một xác suất tương ứng.

Theo Koehn [62], câu ngôn ngữ nguồn $\mathbf{f}$ được tách thành I cụm từ $\overline{f_1}, \overline{f_2}, ..., \overline{f_I}$. Mỗi cụm từ $\overline{f_i}$ trong $\mathbf{f}$ được dịch ra thành một cụm từ $\overline{e_i}$ tương ứng trong $\mathbf{e}$. Quá trình này được thực hiện dựa vào phân phối xác suất $\phi(\overline{f_i}|\overline{e_i})$. Ngoài ra, các cụm từ $\overline{e_i}$ được sắp xếp lại theo một thứ tự nhất định dựa trên mô hình chuyển đổi $d(a_i - b_{i-1})$, với a_i là vị trí bắt đầu của cụm từ $\overline{f_i}$ và b_{i-1} là vị trí kết thúc của cụm từ $\overline{e}_{i-1}$. Khi đó, xác suất dịch $Pr(\mathbf{f}|\mathbf{e})$ được tính theo công thức:

$$Pr(\mathbf{f}|\mathbf{e}) = \prod_{i=1}^{I} \phi(\overline{f_i}|\overline{e_i}) d(a_i - b_{i-1}) \tag{1.8}$$

1.3.3.3 Mô hình dịch dựa trên cú pháp

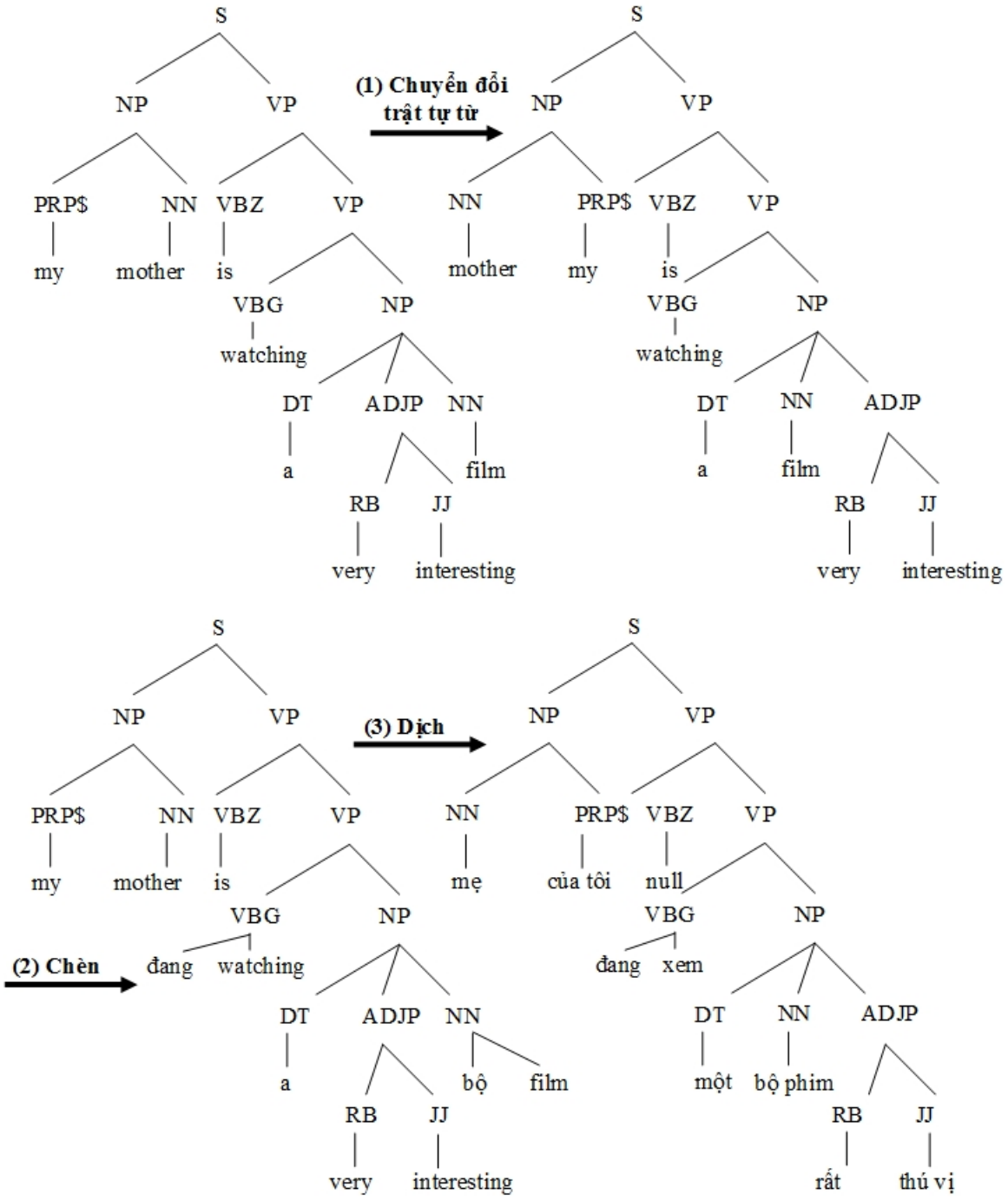
Khác với hai mô hình dịch dựa trên từ và cụm từ như đã trình bày ở trên, mô hình dịch dựa trên cú pháp sử dụng thông tin về cú pháp ngôn ngữ (linguistic syntax). Theo Koehn [60], SMT dựa trên cú pháp có một số ưu điểm: (i) việc chuyển đổi trật tự từ được thực hiện theo cú pháp của ngôn ngữ, (ii) dịch các từ chức năng

tốt hơn (ví dụ như giới từ), (iii) dịch các từ có quan hệ cú pháp tốt hơn (ví dụ, việc dịch động từ có thể phụ thuộc vào chủ ngữ hoặc tân ngữ) và (iv) sử dụng mô hình ngôn ngữ cú pháp (syntactic language model). Các mô hình dịch dựa trên cú pháp rất đa dạng, sử dụng các hình thức và đặc trưng ngữ pháp khác nhau [39]. Một số cách tiếp cận thực hiện phân tích cú pháp cho câu nguồn (tree to string - dịch từ cây cú pháp sang chuỗi), một số khác tạo ra cây cú pháp khi sinh ra câu đích (string to tree - dịch từ chuỗi sang cây cú pháp) và một số kết hợp cả hai (tree to tree - dịch từ cây cú pháp sang cây cú pháp).

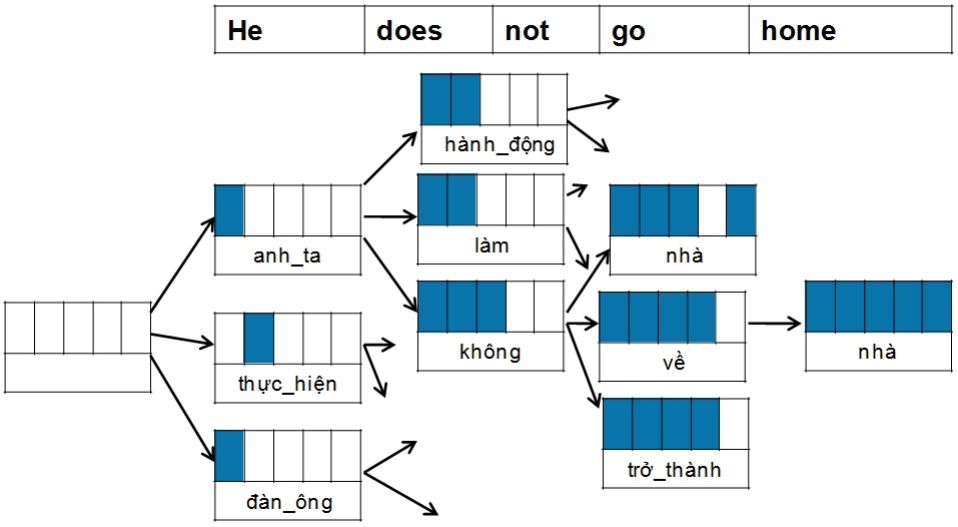
Cách tiếp cận dịch từ cây cú pháp sang chuỗi [46, 71] giả định rằng cú pháp của ngôn ngữ nguồn được biết. Vì thế, cách tiếp cận này có thể được áp dụng khi bộ phân tích cú pháp (parser) của ngôn ngữ nguồn có sẵn. Trong khi đó, các cách tiếp cận dịch từ chuỗi sang cây cú pháp [120, 121] tập trung vào mô hình cú pháp của ngôn ngữ đích trong trường hợp nó có các nguồn tài nguyên cú pháp như ngân hàng câu được chú giải cú pháp (treebank) và bộ phân tích cú pháp. Với cách tiếp cận dịch từ chuỗi sang cây cú pháp như đề xuất của Yamada và Knight [120, 121], câu ngôn ngữ nguồn $\mathbf{f}$ sẽ được phân tích thành cây cú pháp. Cây cú pháp này sẽ được sắp xếp lại để phù hợp với cú pháp của câu ngôn ngữ đích. Sau đó, một số từ mới có thể được chèn vào cây hiện tại cho phù hợp hơn với cú pháp của ngôn ngữ đích. Cuối cùng, các từ trong cây cú pháp của câu ngôn ngữ nguồn sẽ được dịch sang ngôn ngữ đích và ta thu được câu ngôn ngữ đích từ cây cú pháp trên. Một số nghiên cứu mở rộng cách tiếp cận này đã được phát triển, dùng cây cấu trúc cụm từ như Zollmann [36, 129] và cây phụ thuộc của Shen [107]. Cách tiếp cận dịch từ cây cú pháp sang cây cú pháp [22, 73] yêu cầu việc phân tích cú pháp được thực hiện ở cả hai ngôn ngữ (nguồn và đích), công việc này đòi hỏi tăng thêm chi phí thực hiện.

Hình 1.7 mô tả các bước làm việc của một mô hình dịch dựa trên cú pháp theo cách tiếp cận dịch từ chuỗi sang cây cú pháp để dịch một câu từ tiếng Anh sang tiếng Việt [88], gồm 3 bước:

1. Chuyển đổi trật tự từ trên cây cú pháp tiếng Anh: *my mother* $\rightarrow$ *mother my*, *a very interesting film* $\rightarrow$ *a film very interesting*. Sau bước chuyển đổi này, kết quả nhận được là cây cú pháp tiếng Anh có trật tự các nút lá gần với trật tự từ trong câu tiếng Việt nhất.



Hình 1.7: Quá trình dịch dựa trên cú pháp theo cách tiếp cận dịch từ chuỗi sang cây cú pháp, gồm 3 bước: (1) chuyển đổi trật tự từ, (2) chèn và (3) dịch.



HÌNH 1.8: Quá trình dịch được thực hiện từ trái sang phải và mở rộng không gian giả thuyết.

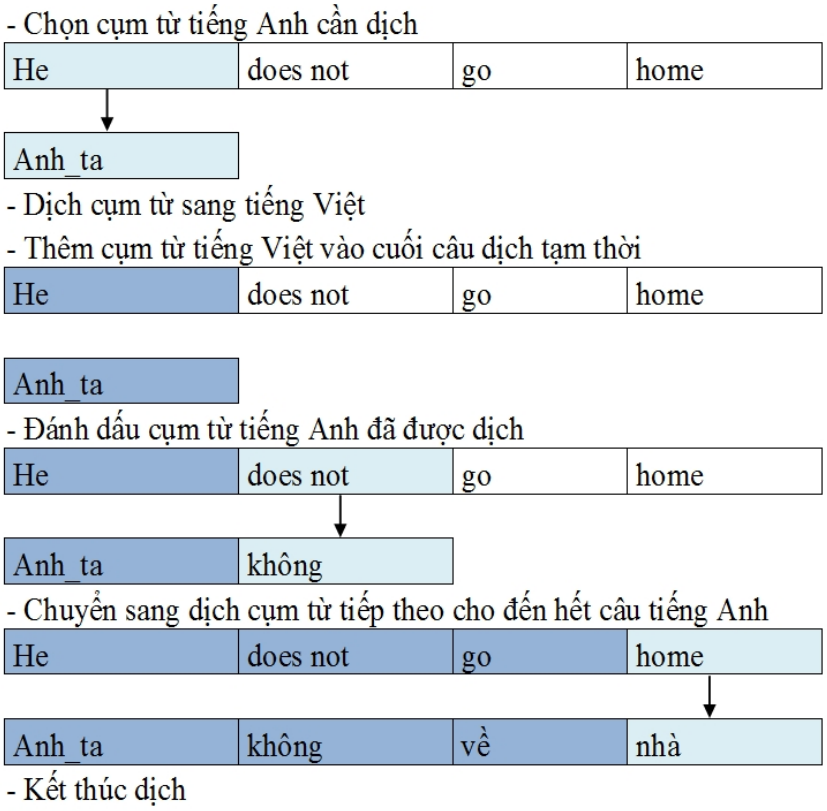
2. Chèn một số nút vào cây cú pháp: *đang*, *bộ*. Các nút được chèn là các nút tiếng Việt, vì vậy thao tác chèn giúp cho câu dịch tiếng Việt được trôi chảy và tự nhiên hơn.

3. Dịch các nút là từ tiếng Anh sang tiếng Việt: *mother* $\rightarrow$ *mẹ*, *my* $\rightarrow$ *của tôi*, *watching* $\rightarrow$ *xem*, *a* $\rightarrow$ *một*, *film* $\rightarrow$ *bộ phim*, *very* $\rightarrow$ *rất*, *interesting* $\rightarrow$ *thú vị*. Các từ vừa được chèn ở bước 2 được giữ nguyên.

1.3.4 Giải mã

Ở trên, chúng tôi đã trình bày về hai trong ba thành phần của một hệ thống SMT, đó là mô hình dịch và mô hình ngôn ngữ. Thành phần còn lại là bộ giải mã (decoder). Nhiệm vụ của bộ giải mã là tìm câu $\mathbf{e}$ ở ngôn ngữ đích sao cho tích $Pr(\mathbf{f}|\mathbf{e})Pr(\mathbf{e})$ trong công thức (1.2) đạt giá trị cực đại với mỗi câu đầu vào $\mathbf{f}$ ở ngôn ngữ nguồn. Theo Koehn [60], các mô hình xác suất trong SMT gán điểm số cho tất cả các bản dịch có thể của một câu đầu vào ở ngôn ngữ nguồn (câu nguồn).

Trong quá trình giải mã, chúng ta xây dựng bản dịch theo từng từ một, từ đầu đến cuối. Các mô hình dựa trên từ và dựa trên cụm từ phù hợp với điều này, vì nó cho phép tính toán điểm số cho các bản dịch một phần (partial translation). Trước khi dịch một câu đầu vào ở ngôn ngữ nguồn, đầu tiên chúng ta tham khảo



HÌNH 1.9: Minh họa quá trình giải mã câu đầu vào $\mathbf{f}$ = "*He does not go home*" từ tiếng Anh sang tiếng Việt.

bảng dịch và tìm kiếm các lựa chọn dịch thích hợp. Trong quá trình giải mã, chúng ta lưu trữ các bản dịch một phần trong một cấu trúc dữ liệu được gọi là giả thuyết (hypothesis). Bộ giải mã đưa ra các hình thức mở rộng những giả thuyết này bằng cách quyết định cụm từ dịch tiếp theo, như mô tả ở Hình 1.8. Do sự phức tạp tính toán của giải mã (NP-đầy đủ), chúng ta cần phải hạn chế không gian tìm kiếm. Chúng ta làm điều này bằng cách tái tổ hợp, dùng kỹ thuật quy hoạch động để loại bỏ giả thuyết mà có thể không phải là một phần của bản dịch tốt nhất. Giới hạn về sắp xếp lại (trật tự từ) cũng làm giảm đáng kể không gian tìm kiếm. Do không gian tìm kiếm là rất lớn, nên bộ giải mã trong mô hình SMT thường áp dụng các thuật toán tìm kiếm tối ưu. Thuật toán mà bộ giải mã thường áp dụng là A*, một kỹ thuật tìm kiếm chuẩn trong trí tuệ nhân tạo [60]. Thuật toán A* có thể tóm tắt như sau: tại mỗi bước mở rộng không gian tìm kiếm, ta sử dụng các hàm ước lượng, đánh giá trọng số để kết quả tìm được luôn là tốt nhất có thể và là kết quả tìm thấy đầu tiên. Hình 1.9 minh họa quá trình giải mã câu đầu vào $\mathbf{f}$ = "*He does not go home*" từ tiếng Anh sang tiếng Việt.

1.3.5 Đánh giá chất lượng dịch

Một chủ đề được tranh luận sôi nổi trong MT là làm thế nào để đánh giá chất lượng dịch, bởi vì có nhiều bản dịch hợp lệ cho mỗi câu đầu vào [60]. Như vậy, chúng ta cần một (hoặc một số) cách định lượng để đánh giá chất lượng hệ thống MT hoặc ít nhất là một cách để có thể biết một hệ thống tốt hơn hệ thống khác hoặc nếu có sự thay đổi trong hệ thống dẫn đến một sự cải tiến. Để đánh giá độ chính xác của bản dịch, chúng ta có thể đánh giá thủ công bởi con người hoặc đánh giá tự động bằng máy tính.

Phương án đánh giá bản dịch bởi con người tuy dễ thực hiện nhưng chi phí rất lớn. Trong trường hợp bản dịch có kích thước càng lớn thì phương pháp này càng kém hiệu quả. Ngày nay, các mô hình MT đều áp dụng phương pháp đánh giá tự động, chi phí thấp nhưng hiệu quả khá cao. Có một số phương pháp đánh giá tự động chất lượng dịch như BLEU[7] [93], NIST[8] [31] và TER[9] [109]. Trong đó, phương pháp đánh giá tự động phổ biến nhất là phương pháp BLEU. Phương pháp này được đề xuất bởi IBM tại hội nghị ACL ở Philadelphia vào tháng 7-2002 [93]. Ý tưởng chính của phương pháp này là so sánh kết quả bản dịch tự động bằng máy với các bản dịch mẫu của con người, bản MT nào càng giống với bản dịch mẫu của con người thì bản dịch đó càng chính xác.

Việc so sánh được thực hiện dựa vào kết quả thống kê sự trùng khớp của các *n-gram* trong hai bản dịch có tính đến thứ tự của chúng trong câu. Giả sử chúng ta có hai bản MT (tiếng Anh) của một câu nguồn tiếng Việt như sau:

- Bản MT 1: *It is a guide to action which ensures that the military always obeys the commands of the party.*
- Bản MT 2: *It is to insure the troops forever hearing the activity guidebook that party direct.*

Chúng ta so sánh với ba bản dịch mẫu:

- Bản dịch mẫu 1: *It is a guide to action that ensures that the military will forever heed Party commands.*

[7]Bilingual Evaluation Understudy
[8]National Institute of Standards and Technology
[9]Translation Error Rate

- Bản dịch mẫu 2: *It is the guiding principle which guarantees the military forces always being under the command of the Party.*
- Bản dịch mẫu 3: *It is the practical guide for the army always to heed the directions of the party.*

Có thể thấy rằng, bản MT thứ nhất có nhiều từ và cụm từ chung với các bản dịch mẫu hơn bản MT thứ hai. Như vậy, chúng ta có thể kết luận bản MT thứ nhất chính xác hơn bản MT thứ hai.

Tổng quát, với bản MT C và bản dịch mẫu R, phương pháp BLEU trước hết thống kê số lần tối thiểu các cụm *n-gram* xuất hiện trong từng cặp câu, sau đó chia cho tổng số cụm *n-gram* trong C. Tỷ lệ trùng khớp p_n của C và R được tính theo công thức:

$$p_n = \frac{\sum_{c \in C} \sum_{n-gram \in c} Count_{clip}(n-gram)}{\sum_{c' \in C} \sum_{n-gram' \in c'} Count_{clip}(n-gram')} \tag{1.9}$$

Trong đó, $Count_{clip}(n-gram)$ là số lượng tối thiểu cụm $n-gram$ có trong R và $Count_{clip}(n-gram')$ là số lượng cụm $n-gram'$ có trong C.

Điểm BLEU đánh giá bản MT C với bản dịch mẫu R được tính theo công thức (1.10). Trong đó, w_n và N lần lượt là trọng số (tổng các trọng số w_n bằng 1) và độ dài (tính theo đơn vị từ) các $n-gram$ được sử dụng:

$$BLEU = BP * \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{1.10}$$

Với giá trị BP được tính theo công thức (1.11):

$$BP = \begin{cases} 1 & \text{nếu } c > r \\ e^{1-r/c} & \text{nếu } c \leq r \end{cases} \tag{1.11}$$

Ở đây, c là độ dài của bản MT và r là độ dài của bản dịch mẫu. Giá trị BLEU đánh giá mức độ tương ứng giữa hai bản dịch. Bản dịch nào có điểm BLEU càng cao, chứng tỏ độ trùng khớp giữa bản MT và bản dịch mẫu càng nhiều, thì bản dịch đó càng chính xác.

Đối với phương pháp NIST [31], việc chọn lựa các *n-gram* và thông tin trên mỗi *n-gram* sẽ được sử dụng để phục vụ việc đánh giá. Sự biến đổi có thể của

điểm đánh giá trên một *n-gram* nếu chúng ta thay đổi vị trí các phần tử trên cùng một *n-gram* cho chúng ta thấy rằng điểm số cũng sẽ thay đổi nếu chúng ta thay đổi vị trí của các *n-gram* trên cùng một phân đoạn [43].

1.4 Thảo luận

Dịch máy dựa trên phương pháp thống kê đang là một hướng phát triển đầy tiềm năng bởi những ưu điểm vượt trội so với các phương pháp khác. Nói chung, hiệu quả (chất lượng dịch) của một hệ thống SMT tỷ lệ thuận với số lượng và chất lượng của ngữ liệu song ngữ được sử dụng để xây dựng hệ thống dịch. Tuy nhiên, ngữ liệu song ngữ hiện vẫn còn hạn chế cả về kích thước lẫn chất lượng. Ngoài ra, việc phát triển các phương pháp để làm tăng chất lượng dịch dựa trên ngữ liệu hiện có đang là một vấn đề mở. Hiện nay, các nghiên cứu để làm tăng chất lượng dịch vẫn đang được tiến hành phù hợp với đặc điểm của các cặp ngôn ngữ [95, 104, 123]. Trong số đó có nghiên cứu của tác giả Đinh Điền [30, 48] về khai phá song ngữ phục vụ cho dịch máy Anh - Việt. Ở đây, tác giả đã khai phá bằng cách học từ song ngữ để rút ra các luật chuyển đổi (transfer rule) để phục vụ cho việc chuyển ngữ từ tiếng Anh sang tiếng Việt. Khác với [30, 48], trong luận án này chúng tôi huấn luyện bằng mô hình IBM cải tiến.

Từ những phân tích, đánh giá các nghiên cứu liên quan ở trên, chúng tôi nhận thấy một số vấn đề còn tồn tại, cụ thể như sau:

- Thứ nhất, đối với bài toán xây dựng ngữ liệu cho SMT, chúng ta có thể khai thác từ hai nguồn: Web và sách điện tử song ngữ. Đối với nguồn từ Web, các nghiên cứu trước chưa kết hợp giữa các đặc trưng dựa trên nội dung với các đặc trưng dựa cấu trúc của trang *web* để làm tăng độ chính xác trong việc rút trích các văn bản song ngữ. Ngoài ra, sách điện tử song ngữ chứa một số lượng lớn các văn bản song ngữ được dịch cẩn thận. Đây là nguồn dữ liệu rất tiềm năng để bổ sung ngữ liệu song ngữ cho SMT. Tuy nhiên, theo hiểu biết của chúng tôi, hiện chưa có nghiên cứu nào khai thác nguồn này (sách điện tử song ngữ) cho cặp ngôn ngữ Anh - Việt.
- Thứ hai, gióng hàng từ đóng vai trò rất quan trọng cho sự thành công của một hệ thống SMT. Như chúng tôi đã phân tích ở trên, sử dụng các mô hình thống kê (các mô hình IBM) để gióng hàng từ cho SMT có nhiều ưu điểm so

với cách tiếp cận dựa trên mô hình phân biệt. Tuy nhiên, nếu chỉ sử dụng các mô hình thống kê thuần túy thì rất khó để đạt được gióng hàng với độ chính xác cao. Vì vậy, sử dụng thêm các nguồn tri thức bên ngoài như thông tin về từ vựng, thông tin về cú pháp là thật sự cần thiết để cải thiện chất lượng của gióng hàng.

- Thứ ba, các cụm từ song ngữ được sử dụng để bổ sung nguồn tri thức song ngữ cho các hệ thống SMT. Bouamor và cộng sự [9] đã chỉ ra rằng, các cụm từ song ngữ được sử dụng để cải thiện chất lượng dịch cho SMT. Trong các nghiên cứu liên quan sử dụng mẫu cú pháp để xác định cụm từ song ngữ. Việc so khớp các mẫu cú pháp được thực hiện ở cả câu nguồn và câu đích. Cách làm này chỉ rút trích được các cụm từ song ngữ với số lượng hạn chế.

Trong các chương tiếp theo của luận án, chúng tôi sẽ tập trung giải quyết các tồn tại đã nêu thông qua ba bài toán tương ứng với ba chương trong luận án (từ Chương 2 đến Chương 4):

1. Xây dựng ngữ liệu song ngữ cho SMT;
2. Gióng hàng từ cho SMT;
3. Xác định cụm từ song ngữ cho SMT.

Đối với bài toán thứ nhất, chúng tôi khai thác từ hai nguồn: Web và sách điện tử song ngữ. Đối với nguồn từ Web, chúng tôi rút trích các văn bản song ngữ từ các trang *web* song ngữ Anh - Việt; đưa ra hai phương pháp thiết kế các đặc trưng dựa trên nội dung: dựa trên *cognate* và dựa trên việc xác định các phân đoạn dịch. Sau đó, chúng tôi kết hợp các đặc trưng dựa trên nội dung với các đặc trưng dựa trên cấu trúc và mô hình hóa bài toán này như bài toán phân loại để trích rút các văn bản song ngữ. Đối với nguồn từ sách điện tử song ngữ, chúng tôi sử dụng dữ liệu ngôn ngữ thông qua một hệ thống SMT để rút trích các cặp câu song ngữ Anh - Việt (thông qua việc gióng hàng đoạn/câu).

Với bài toán thứ hai, chúng tôi đề xuất một số cải tiến đối với mô hình IBM 1 theo cách tiếp cận dựa trên ràng buộc, bao gồm: ràng buộc neo, ràng buộc về vị trí của từ, ràng buộc về từ loại và ràng buộc về cụm từ. Với mỗi ràng buộc, chúng tôi đưa ra phương pháp tổng quát để tích hợp nó vào thuật toán EM trong quá trình ước lượng tham số của mô hình. Ngoài ra, chúng tôi đưa ra một phương pháp để kết hợp các ràng buộc.

Bài toán thứ ba, chúng tôi đề xuất phương pháp xác định cụm từ song ngữ cho SMT. Chúng tôi sử dụng các mẫu cú pháp kết hợp với gióng hàng cụm từ để xác định cụm từ song ngữ. Các cụm từ song ngữ này được ứng dụng vào việc nâng cao chất lượng dịch cho SMT.

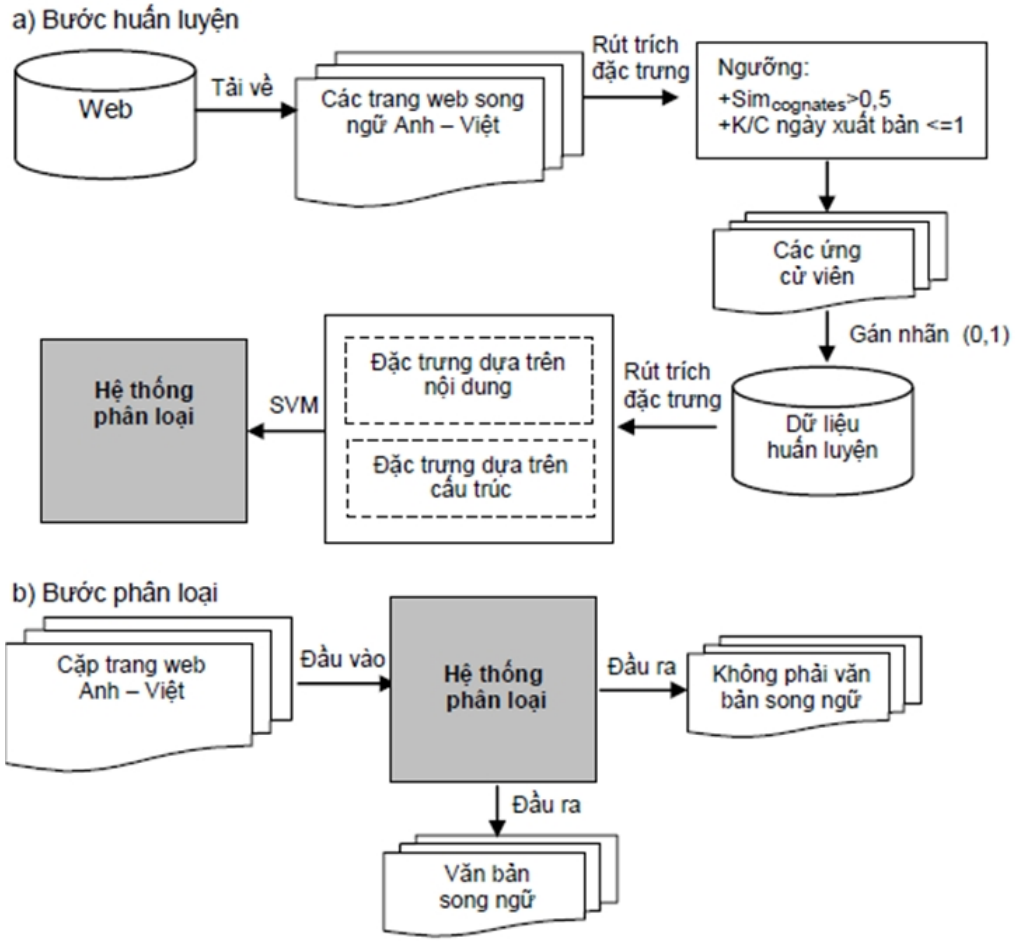
Chương 2

Xây dựng ngữ liệu song ngữ cho dịch máy thống kê

Chương này trình bày về việc xây dựng ngữ liệu song ngữ cho SMT. Để thực hiện công việc này, chúng tôi khai thác từ hai nguồn: Web và sách điện tử (ebook) song ngữ. Đối với nguồn từ Web, chúng tôi tập trung vào rút trích các văn bản song ngữ từ các *web-site* song ngữ Anh - Việt. Trước hết, chúng tôi đề xuất hai phương pháp thiết kế các đặc trưng dựa trên nội dung: (i) sử dụng *cognate* và (ii) sử dụng các phân đoạn dịch (translation segment). Sau đó, chúng tôi kết hợp các đặc trưng dựa trên nội dung với các đặc trưng dựa trên cấu trúc của trang *web* để rút trích các văn bản song ngữ, bằng cách sử dụng phương pháp học máy. Đối với nguồn từ sách điện tử song ngữ, mục tiêu của chúng tôi là rút trích các câu song ngữ. Để thực hiện công việc này, chúng tôi đề xuất phương pháp dựa trên nội dung sử dụng một số mẫu liên kết giữa các khối văn bản trong hai ngôn ngữ. Quá trình thực hiện gồm ba bước chính: tiền xử lý, gióng hàng đoạn và gióng hàng câu. Các thực nghiệm, đánh giá được trình bày ở cuối chương.

2.1 Rút trích văn bản song ngữ từ Web

Với sự bùng nổ của Internet hiện nay, dữ liệu song ngữ ngày càng được mở rộng nhanh chóng. Do vậy, yêu cầu đặt ra là làm thế nào để khai thác tự động nguồn dữ liệu song ngữ này. Muốn khai thác được dữ liệu này để xây dựng ngữ liệu song ngữ, nhiệm vụ đầu tiên và quan trọng nhất là chúng ta cần tìm ra những trang



Hình 2.1: Sơ đồ của hệ thống rút trích văn bản song ngữ từ Web.

web song ngữ. Ở đây, chúng tôi tập trung giải quyết vấn đề này, cụ thể chúng tôi thực hiện rút trích văn bản song ngữ từ Web.

Chúng tôi sử dụng kết hợp giữa các đặc trưng dựa trên nội dung và các đặc trưng dựa trên cấu trúc của các tài liệu HTML (trang *web*) để rút trích các văn bản song ngữ từ Web, bằng cách sử dụng phương pháp học máy [101]. Thuật toán học máy được sử dụng ở đây là máy véc-tơ hỗ trợ (Support Vector Machine - SVM). Hình 2.1 mô tả sơ đồ tổng quát của phương pháp chúng tôi đưa ra. Quá trình thực hiện bao gồm các bước như sau:

1. Sử dụng *crawler* (chương trình tự động duyệt và tải về các trang *web*) từ các *domain* đã được xác định để thu thập các trang *web* ở hai ngôn ngữ, các trang *web* này được gọi là dữ liệu thô (raw data).

2. Từ tập dữ liệu thô thu được ở bước 1, chúng tôi tạo ra các ứng viên của các trang *web* song ngữ dựa vào một số đặc trưng[1] ($sim_{cognate} > 0,5$ và $distance_{date} \leq 1$).

3. Gán nhãn cho các ứng viên này, từ đó xây dựng dữ liệu huấn luyện (training data). Điều này có nghĩa là, chúng tôi đạt được một số cặp trang *web* ở hai ngôn ngữ. Một số cặp trang *web* là song ngữ được gán nhãn 1 và các cặp khác được gán nhãn 0 (chi tiết công việc này chúng tôi sẽ trình bày ở phần thực nghiệm của chương).

4. Thiết kế các đặc trưng dựa trên nội dung và cấu trúc của trang *web*. Từ đó, mỗi trang *web* có thể được biểu diễn bởi một véc-tơ đặc trưng. Biểu diễn này sẽ được sử dụng trong mô hình phân loại.

5. Cuối cùng, sử dụng thuật toán SVM để huấn luyện một hệ thống phân loại dựa trên dữ liệu này. Tức là, nếu chúng ta có một cặp trang *web* ở hai ngôn ngữ, hệ thống phân loại sẽ quyết định cặp này có phải là song ngữ hay không.

2.1.1 Thu thập dữ liệu

Để thực hiện việc thu thập các tài liệu HTML từ Web, chúng tôi sử dụng công cụ Teleport-Pro[2] (như là Web crawler[3]). Teleport-Pro là công cụ được thiết kế để tải về các tài liệu trên mạng thông qua các giao thức HTTP, FTP và thực hiện việc lưu trữ các dữ liệu thu thập được [76]. Ở đây, chúng tôi chọn các URL[4] từ ba *web-site*: BBC, VietnamPlus và VOA News. Ví dụ, URL của trang BBC cho tiếng Anh là "http://www.bbc.co.uk" và "http://www.bbc.co.uk/vietnamese/" cho tiếng Việt. Sau đó, chúng tôi dùng Teleport-Pro để tải về các trang *web* ở hai ngôn ngữ.

2.1.2 Thiết kế các đặc trưng dựa vào nội dung

Để tìm các trang *web* song ngữ, nếu chỉ dựa trên cấu trúc thì chúng ta sẽ không thể thực hiện được trong hai trường hợp sau:

[1] Chúng tôi sẽ làm rõ các đặc trưng này ở các phần tiếp theo.
[2] http://www.tenmax.com/teleport/pro/home.htm
[3] Theo Cambazoglu [14], Web crawling là quá trình tìm kiếm, thu thập và lưu trữ các tài liệu từ Web. Các chương trình máy tính thực hiện công việc này được gọi là Web crawler hoặc spider.
[4] Uniform Resource Locator

1. Hai trang *web* có cấu trúc khác nhau nhưng nội dung giống nhau.

2. Hai trang *web* có cấu trúc giống nhau nhưng nội dung khác nhau.

Hơn nữa, nếu chỉ dựa trên cấu trúc thì chúng ta chưa tận dụng được nội dung của trang *web*, vì nội dung mới là yếu tố quan trọng nhất. Lưu ý rằng, bài toán chúng ta đang giải quyết là tìm hai trang *web* có nội dung giống nhau. Ở đây, chúng ta sẽ không quan tâm đến việc trình bày trang *web* thế nào mà chỉ quan tâm đến nội dung bên trong nó.

Trở lại vấn đề, các trang *web* thu thập được từ các URL ở bước trước được chuyển đổi sang định dạng văn bản. Lưu ý rằng, các trang *web* ban đầu thường chứa các thành phần giao diện người dùng như JavaScript, Flash, vv. Vì vậy, chúng tôi sẽ loại bỏ những thành phần này và chỉ lấy phần văn bản là nội dung chính ở trong trang *web*. Dưới đây, chúng tôi trình bày hai phương pháp thiết kế các đặc trưng dựa vào nội dung: (i) sử dụng *cognate* và (ii) sử dụng các phân đoạn dịch.

2.1.2.1 Sử dụng cognate

Phương pháp này sử dụng các từ cùng nguồn gốc (cognate) hay còn gọi là các từ bất biến giữa hai ngôn ngữ. Các nghiên cứu trước đây dùng *cognate* cho một số bài toán khác nhau, ví dụ như Kondrak [63] sử dụng *cognate* để cải tiến mô hình dịch thống kê hay để gióng hàng từ như trong [111]. Khác với những nghiên cứu này, ở đây chúng tôi sử dụng *cognate* để xác định các văn bản song ngữ. Theo quan sát của chúng tôi, các tài liệu nói chung thường có chứa một số từ cùng nguồn gốc và nếu hai tài liệu là bản dịch lẫn nhau thì các từ cùng nguồn gốc thường giống nhau. Các từ cùng nguồn gốc là những từ được đánh vần tương tự trong hai ngôn ngữ hoặc những từ mà chỉ đơn giản là không được dịch[5] (ví dụ, chữ viết tắt).

Bây giờ, chúng ta có thể thiết kế một đặc trưng để đo độ tương tự về nội dung dựa trên các từ cùng nguồn gốc. Giá trị của nó được tính bằng tỷ lệ giữa số lượng từ cùng nguồn gốc giống nhau ở hai văn bản với số lượng từ cùng nguồn gốc trong một văn bản (ví dụ, văn bản tiếng Anh). Với một cặp văn bản ($Etext$, $Vtext$), trong đó: $Etext$ là viết tắt của văn bản tiếng Anh và $Vtext$ là viết tắt của văn bản tiếng Việt, chúng tôi xác định các tập T_E và T_V chứa các *cognate* ở trong $Etext$ và $Vtext$. Bảng 2.1 là một ví dụ về hai văn bản có chứa các *cognate* tương ứng giữa tiếng Anh và tiếng Việt. Duyệt các văn bản này, chúng tôi đạt

[5]Để tăng số lượng *cognate*, chúng tôi sử dụng thêm thông tin về tên của các nước.

Bảng 2.1: Ví dụ về hai văn bản có chứa các *cognate* tương ứng giữa tiếng Anh và tiếng Việt (các từ in nghiêng).

Văn bản tiếng Anh	**Văn bản tiếng Việt**
Vietnam and *Italy* through three cooperation programmes beginning in *1998* have so far signed more than *60* projects on joint scientific research. Of the figure, *40* projects have been carried out and brought good results.	Từ *1998*, đến nay, *Việt Nam* và *Italy* đã ký kết hơn *60* dự án hợp tác nghiên cứu chung, có khoảng *40* dự án đã được triển khai thực hiện và đạt được kết quả tích cực.
The projects concentrated on the application of *Italian* technology into *Vietnam*, the research of scientific-technological issues of *Vietnam*'s concern, trainee exchange, and provision of research facilities.	Nội dung của các dự án bao gồm: Hợp tác cùng nghiên cứu, tích hợp công nghệ của *Italy* vào *Việt Nam*, cùng nghiên cứu và hoàn thiện trình diễn một số vấn đề khoa học và công nghệ mà phía *Việt Nam* quan tâm; trao đổi thực tập sinh *Italy* và hỗ trợ một số trang thiết bị nghiên cứu cho *Việt Nam*.
Most of the projects make use of *Italy*'s scientific-technological fortes such as restoration of ancient relic sites, medicine, energy and environment, biotechnology and agriculture, basic research and information technology.	Một số dự án đã được triển khai trên cơ sở các thế mạnh khoa học và công nghệ của *Italy*, cũng như những vấn đề khoa học và công nghệ mà Việt Nam quan tâm như bảo tồn phục chế các di tích cổ; y dược và sức khỏe con người.

được các tập T_E={"Vietnam", "Italy", "1998", "60", "40",...} và T_V ={"1998", "Vietnam", "Italy", "60", "40",...}. Độ tương tự về *cognate* của $Vtext$ với $Etext$ được xác định theo công thức (2.1)[6]. Nếu $sim_{cognate}(Etext, Vtext)$ lớn hơn ngưỡng cho trước thì hai văn bản ($Etext$,$Vtext$) là một cặp ứng viên.

$$sim_{cognate}(Etext, Vtext) = \frac{|T_E \cap T_V|}{|T_E|} \tag{2.1}$$

Ngoài việc sử dụng thông tin *cognate*, chúng tôi nhận thấy rằng độ dài văn bản và số lượng các đoạn cũng cung cấp chứng cứ để đo độ tương tự về nội dung giữa hai văn bản. Văn bản song ngữ thường tương đồng về độ dài văn bản và số lượng đoạn. Vì vậy, với một cặp văn bản chúng tôi thiết kế ba đặc trưng như sau:

- Độ tương tự về *cognate* theo công thức (2.1);
- Tỷ lệ về độ dài văn bản tính theo đơn vị từ;

[6]Lưu ý, theo cách tính của chúng tôi $sim_{cognate}(Etext, Vtext) \neq sim_{cognate}(Vtext, Etext)$

- Tỷ lệ về số đoạn giữa hai văn bản.

2.1.2.2 Sử dụng các phân đoạn dịch

Phương pháp này dựa trên việc xác định các phân đoạn dịch, như là sự mở rộng định nghĩa về bản văn song ngữ để so khớp các tài liệu, đoạn, câu và từ. Điều này sẽ giúp chúng ta rút trích các khối dịch thích hợp trong các trang *web* song ngữ. Để phục vụ việc xây dựng ngữ liệu cho SMT, ở đây chúng tôi tập trung vào việc xác định các đoạn dịch.

Giả sử chúng ta đang xem xét hai trang *web*, trong đó mỗi trang có mối quan hệ dịch với trang kia. Theo quan sát của chúng tôi, có ba loại quan hệ giữa các trang *web* song ngữ, bao gồm:

1. Trang ở ngôn ngữ đích là bản dịch đầy đủ của trang gốc ở ngôn ngữ nguồn. Với loại này, thông thường bản dịch được thực hiện ở cấp độ câu.

2. Trang ở ngôn ngữ đích được tạo ra bằng cách dịch một số phần từ trang gốc. Trong các phần này, các văn bản được dịch chặt chẽ. Các phần còn lại được thay đổi nhiều ở trong ngôn ngữ đích, chỉ giữ lại ý tưởng chính. Chúng tôi nhận thấy rằng các phần được dịch thường là đoạn trong một số trang *web* song ngữ giữa tiếng Anh và tiếng Việt.

3. Trong loại này, người dịch chỉ giữ lại một số thông tin từ trang gốc, hoặc thậm chí kết hợp thông tin từ các trang khác nhau để tạo ra một trang mới trong ngôn ngữ khác.

Giả sử chúng ta đang làm việc với hai ngôn ngữ là tiếng Anh và tiếng Việt. Theo quan điểm này, chúng tôi biểu diễn nội dung của một trang *web* (body text) bằng một chuỗi các đoạn. Ký hiệu $Epage$, $Etext$, $Vpage$ và $Vtext$ lần lượt là trang *web* tiếng Anh, nội dung của trang *web* tiếng Anh, trang *web* tiếng Việt, nội dung của trang *web* tiếng Việt. Khi đó, $Etext$ được biểu diễn như là một chuỗi các đoạn $pe_1 pe_2 \dots pe_n$ và $Vtext$ được biểu diễn như là một chuỗi các đoạn $pv_1 pv_2 \dots pv_m$. Trong đó, pe_i và pv_j tương ứng là các đoạn trong văn bản tiếng Anh và tiếng Việt.

Các nghiên cứu liên quan trong [16, 24, 76] khai thác văn bản song ngữ từ Web. Họ so sánh toàn bộ nội dung của hai trang *web* để xác định xem nó là song ngữ hay không. Theo quan điểm của chúng tôi, nếu hai trang *web* có chứa một số

đoạn song ngữ thì những đoạn này cần được rút trích. Điều này chỉ đòi hỏi đánh giá một phần của hai trang, như vậy tránh được nhược điểm từ các nghiên cứu trước đó là nó sẽ trả về điểm số[7] thấp ngay cả hai trang chứa các đoạn song ngữ. Khác với cách tiếp cận trong [84, 85] nhằm mục đích để rút trích các câu song ngữ, ở đây chúng tôi sử dụng thông tin về các đoạn dịch để xác định hai trang *web* song ngữ. Đối với mức độ đoạn, chúng ta có thể thích ứng với các trường hợp dịch *1-n* hay *n-1* giữa các câu trong hai ngôn ngữ, điều này cũng phù hợp với quan sát của chúng tôi. Lưu ý rằng các đoạn có thể được xác định dễ dàng từ các trang *web* dựa trên thẻ $< p >$.

Bây giờ chúng ta phải tìm các đoạn trong một ngôn ngữ là bản dịch của một ngôn ngữ khác. Thông tin này sau đó được sử dụng để xác định xem hai trang *web* có phải là song ngữ hay không. Chúng tôi thiết kế hàm $Similarity_{paragraph}(pe, pv)$ để đo mối quan hệ dịch giữa pe và pv. Điều này có thể được tính toán dựa trên dịch từ vựng dùng từ điển song ngữ hoặc bằng các phương pháp khác. Như vậy, đối với mỗi pe_i chúng ta cần tìm pv_j thích hợp nhất được ký hiệu như trong công thức (2.2).

$$pv_j = \arg\max_{pv_k} Similarity_{paragraph}(pe_k, pv_i),\ \mathrm{k} = 1, \ldots, \mathrm{n} \tag{2.2}$$

Từ cặp đoạn dịch này, chúng ta có thể đánh giá mức độ về mối quan hệ giữa $Etext$ và $Vtext$. Và sau đó nó sẽ được biểu diễn như một tập đặc trưng và trọng số của nó trong mô hình học được sử dụng để xác định các trang song ngữ. Phần tiếp theo sẽ mô tả chi tiết các đặc trưng này (chúng tôi gọi nó là đặc trưng dựa vào nội dung) và chỉ ra cách làm thế nào để tính toán nó.

Các nghiên cứu trước đây thường sử dụng dịch từ vựng nhận được từ một từ điển song ngữ để đo độ tương tự về nội dung của hai văn bản, chẳng hạn như trong [16, 101]. Cách tiếp cận này có thể phải đối mặt với khó khăn vì một từ thường có nhiều bản dịch của nó. Khác với những nghiên cứu trước, ở đây chúng tôi sử dụng một hệ thống SMT. Với cách làm này, chúng ta có thể tận dụng những lợi thế của SMT như giải quyết vấn đề nhập nhằng, dịch theo cụm từ, trật tự từ.

Công thức (2.2) cho thấy làm thế nào để chọn đoạn dịch tốt nhất. Chúng tôi sử dụng độ đo BLEU [93] để tính độ tương tự giữa bản dịch tiếng Việt của pe_i và pv_j (biểu diễn bởi hàm $Similarity_{paragraph}$). Như vậy, từ hai chuỗi $pe_1 pe_2 \ldots pe_n$

[7]Điểm số ở đây là độ tương tự về nội dung giữa 2 văn bản.

và $pv_1 pv_2 \dots pv_m$ chúng ta có thể dễ dàng tìm thấy một liên kết giữa chúng, trong đó mỗi pe_i gióng hàng với đoạn dịch tốt nhất của nó là pv_j (ngoại trừ các đoạn đã được chọn). Lưu ý rằng, các cặp đoạn với điểm BLEU cao hơn sẽ được lựa chọn trước. Một số đoạn không có đoạn dịch tương ứng sẽ được liên kết với đoạn rỗng (empty paragraph) và độ tương tự giữa nó bằng 0. Từ những cặp đoạn này, chúng tôi chỉ lựa chọn trong đó một số cặp có điểm BLEU lớn hơn một ngưỡng. Những cặp lựa chọn được coi là đoạn song ngữ.

Bây giờ, chúng tôi thiết kế hai đặc trưng dựa trên nội dung, đây được xem như là chứng cứ để xác định liệu cặp ($Etext$, $Vtext$) có phải là một văn bản song ngữ hay không.

1. Giá trị trung bình $Similarity_{paragraph}$ của các cặp đoạn được chọn (ký hiệu là $avgSim_{paragraph}$). Giá trị này cho biết mức độ dịch của các phần trong hai văn bản (ở đây là các đoạn) đã được xác định có đủ mối quan hệ dịch. Giá trị cao thể hiện rằng những đoạn song ngữ chặt chẽ.

2. Tỷ lệ giữa số cặp đoạn được lựa chọn (các đoạn song ngữ) và tổng các đoạn của $Etext$ và $Vtext$ (ký hiệu là $rate_{translation}$). Giá trị này gần bằng 1 có nghĩa $Vtext$ là dịch đầy đủ của $Etext$ và bằng 0 nếu ngược lại. Giá trị này nằm giữa 0 và 1 có nghĩa là $Vtext$ là bản dịch một phần của $Etext$.

2.1.3 Thiết kế các đặc trưng dựa vào cấu trúc

Bên cạnh việc tìm các trang *web* song ngữ dựa vào nội dung của văn bản, độ tương tự về cấu trúc của các trang *web* cũng cung cấp thông tin hữu ích để xác định liệu một cặp trang *web* là song ngữ hay không. Phương pháp này sử dụng giả thuyết rằng "các trang *web* song ngữ được trình bày với cấu trúc tương tự nhau". Lưu ý rằng, phương pháp dựa vào cấu trúc không đòi hỏi tri thức về ngôn ngữ. Ở đây, chúng tôi áp dụng phương pháp được trình bày trong [101] để thiết kế các đặc trưng dựa trên cấu trúc.

Quá trình phân tích cấu trúc được thực hiện theo hai bước. Tại bước đầu tiên, hai trang *web* là cặp ứng viên được phân tích thông qua một bộ phân tích thẻ HTML hoạt động như một bộ chuyển đổi, tạo ra một thứ tự tuyến tính có chứa ba loại thẻ:

1. [START:element_label], ví dụ: [START:H3]
2. [END:element_label], ví dụ: [END:H3]
3. [Chunk:length], ví dụ: [Chunk:250]

Ở bước thứ hai, chúng tôi thực hiện gióng hàng các thẻ thu được ở bước 1 bằng cách sử dụng kỹ thuật quy hoạch động. Sau khi gióng hàng, chúng tôi tính toán bốn giá trị vô hướng đặc trưng cho chất lượng của quá trình gióng hàng:

- *dp* Phần trăm khác nhau, chỉ những thành phần được gióng không xuất hiện đồng thời trong hai trang.
- n Số lượng chữ trong thẻ (non-markup text) không bằng nhau.
- r Sự tương quan về độ dài của các thành phần chữ trong thẻ được gióng.
- p Độ tin cậy của hệ số tương quan r.

Ngoài ra, chúng tôi nhận thấy rằng trên các *web-site* tin tức song ngữ, bản dịch của trang gốc sẽ được tạo ra trong một thời gian ngắn sau khi bản gốc được xuất bản. Do đó, sử dụng đặc trưng về ngày xuất bản, chúng ta có thể loại bỏ nhiều cặp không phải bản dịch của nhau. Ví dụ, các *web-site* tin tức song ngữ như BBC, VOA, Việt Nam Plus,... bản dịch tiếng Việt được công bố trong cùng một ngày hoặc sau đó một ngày so với các trang tiếng Anh tương ứng [24]. Để lấy thông tin này, chúng tôi tiến hành phân tích các thẻ HTML và sau đó bổ sung đặc trưng này vào tập đặc trưng dựa trên cấu trúc như đã trình bày ở trên.

2.1.4 Mô hình hóa bài toán phân loại

Từ các kết quả phân tích dựa trên nội dung và cấu trúc của mỗi cặp trang *web* đã trình bày, chúng tôi thu được 10 đặc trưng như tổng hợp ở Bảng 2.2. Các đặc trưng này được chia làm hai nhóm: (i) đặc trưng về cấu trúc (các đặc trưng từ 1 đến 5) và (ii) đặc trưng về nội dung (các đặc trưng từ 6 đến 10). Bây giờ chúng ta có thể dễ dàng để mô hình hóa thành bài toán phân loại (classification problem). Mỗi cặp ứng viên của trang *web* song ngữ được biểu diễn bởi một véc-tơ đặc trưng.

Gọi $F = \{f_1, f_2, ..., f_m\}$ là tập đặc trưng, $D = \{d_1, d_2, ..., d_n\}$ là tập chứa tất cả các cặp ứng viên và $C = \{0, 1\}$ là tập các loại (0: không song ngữ, 1: song

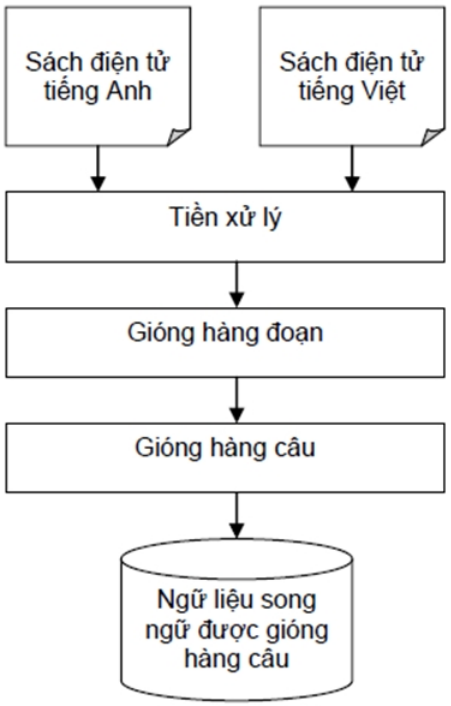
Bảng 2.2: Tổng hợp các đặc trưng.

STT	Ký hiệu	Mô tả
1.	dp	Phần trăm khác nhau, chỉ những thành phần được gióng không xuất hiện đồng thời trong hai trang.
2.	n	Số lượng chữ trong thẻ không bằng nhau.
3.	r	Sự tương quan về độ dài của các thành phần chữ trong thẻ được gióng.
4.	p	Độ tin cậy của hệ số tương quan r.
5.	$distance_{date}$	Khoảng cách về ngày xuất bản giữa hai trang *web*.
6.	$sim_{cognate}$	Độ tương tự về *cognate*.
7.	$rate_{length}$	Tỷ lệ về độ dài văn bản.
8.	$rate_{paragraph}$	Tỷ lệ về số đoạn giữa hai văn bản.
9.	$avgSim_{paragraph}$	Giá trị trung bình $Similarity_{paragraph}$ của các cặp đoạn được chọn.
10.	$rate_{translation}$	Tỷ lệ giữa số cặp đoạn được lựa chọn (các đoạn song ngữ) và tổng các đoạn của hai văn bản.

ngữ). Khi đó, mỗi cặp ứng viên $d_i \in D$ được biểu diễn bởi véc-tơ đặc trưng $d_i = (f_{1i}, f_{2i}, ..., f_{mi})$. Chúng tôi gán nhãn cho chúng là 1 hoặc 0 nếu mỗi cặp tương ứng là song ngữ hoặc không song ngữ. Bằng cách này, chúng ta sẽ có được dữ liệu huấn luyện. Ở đây, chúng tôi sử dụng thuật toán SVM để huấn luyện hệ thống phân loại. Đối với một cặp trang *web* mới, đầu tiên chúng tôi rút trích tập đặc trưng F để có thể biểu diễn nó như là một véc-tơ. Véc-tơ này đi qua hệ thống phân loại và nhận được kết quả là 1 hoặc 0.

2.2 Rút trích câu song ngữ từ sách điện tử

Để xây dựng ngữ liệu song ngữ từ sách điện tử, chúng tôi đề xuất phương pháp dựa trên nội dung, sử dụng một số mẫu liên kết giữa các khối văn bản trong hai ngôn ngữ và kết hợp với một hệ thống SMT để gióng hàng đoạn/câu. Hình 2.2 minh họa sơ đồ tổng quát của phương pháp chúng tôi đề xuất để gióng hàng đoạn/câu cho sách điện tử song ngữ Anh - Việt. Như thể hiện trong hình, phương pháp của chúng tôi bao gồm ba bước chính: tiền xử lý, gióng hàng đoạn và gióng hàng câu. Trong các phần tiếp theo, chúng tôi sẽ mô tả chi tiết các bước này. Ngoài ra, chúng tôi cũng sẽ trình bày cách thức làm thế nào để đo độ tương tự giữa hai khối văn bản. Chúng tôi sẽ giải quyết hai vấn đề sau đây:



HÌNH 2.2: Sơ đồ mô tả quá trình gióng hàng đoạn/câu cho sách điện tử song ngữ Anh - Việt.

1. Gióng hàng đoạn;

2. Gióng hàng câu.

Cho sách điện tử tiếng Anh $\boldsymbol{E}$ chứa I khối (đoạn hoặc câu) $ue_1, ..., ue_I$ và sách điện tử tiếng Việt $\boldsymbol{V}$ chứa J khối $uv_1, ..., uv_J$, chúng tôi định nghĩa một liên kết $l = (i, j)$ tồn tại nếu ue_i là bản dịch (hoặc bản dịch một phần) của uv_j và/hoặc ngược lại. Khi đó, một gióng hàng $\boldsymbol{A}$ (giữa $\boldsymbol{E}$ và $\boldsymbol{V}$) được định nghĩa là một tập hợp con của tập tích Đề-Các của các vị trí đoạn/câu. Một cách hình thức, bài toán gióng hàng đoạn/câu được biểu diễn như trong công thức (2.3).

$$\boldsymbol{A} \subseteq \{(i, j) : i = 1, ..., I; j = 1, ..., J\} \tag{2.3}$$

Nhiệm vụ của chúng tôi là tìm gióng hàng $\boldsymbol{A}$, liên kết các đoạn/câu trong $\boldsymbol{E}$ với các đoạn/câu tương ứng trong $\boldsymbol{V}$. Bảng 2.3 minh họa gióng hàng câu trong một đoạn song ngữ Anh - Việt. Trong ví dụ này, gióng hàng $\boldsymbol{A}$ là $\{(1-1), (2-2), (3-3), (4-4), (5-5), (5-6), (6-7)\}$. Chúng ta có thể thấy rằng, hầu hết các câu tiếng Anh liên kết với chính xác một câu tiếng Việt. Tuy nhiên, trong một số

Bảng 2.3: Ví dụ về gióng hàng câu trong một đoạn văn bản song ngữ Anh - Việt.

Đoạn văn bản tiếng Anh	**Đoạn văn bản tiếng Việt**
1. I had known him since 1984, when he came to Manhattan to have lunch with Time's editors and extol his new Macintosh.	**1.** Tôi quen ông từ năm 1984, khi ông đến Manhattan để ăn trưa cùng với những biên tập viên của tạp chí Time và nhân tiện giới thiệu luôn chiếc máy Macintosh (Mac) mới của mình.
2. He was petulant even then, attacking a Time correspondent for having wounded him with a story that was too revealing.	**2.** Thậm chí lúc đó ông đã nổi nóng, và tấn công một phóng viên của tạp chí Time vì đã làm ông tổn thương bằng một câu chuyện quá lố.
3. But talking to him afterward, I found myself rather captiv ated, as so many others have been over the years, by his engaging intensity.	**3.** Nhưng sau này khi có cơ hội nói chuyện với Jobs, tôi thấy mình bị cuốn hút, giống như bao người khác trong nhiều năm qua, bởi sự hấp dẫn tuyệt vời toát lên từ con người ông.
4. We stayed in touch, even after he was ousted from Apple.	**4.** Chúng tôi giữ liên lạc, kể cả khi ông không còn làm ở Apple nữa.
5. When he had something to pitch, such as a NeXT computer or Pixar movie, the beam of his charm would suddenly refocus on me, and he would take me to a sushi restaurant in Lower Manhattan to tell me that whatever he was touting was the best thing he had ever produced.	**5.** Khi có một cái gì đó muốn khoe, ví dụ như một chiếc máy tính của NeXT hay một bộ phim của Pixar, ông đều chia sẻ với tôi những điều tuyệt vời đó. **6.** Ông mời tôi đến một nhà hàng sushi ở Hạ Manhattan và nói với tôi rằng bất cứ những gì ông đang đưa ra thị trường đều là những thứ tốt nhất mà ông đã tạo ra.
6. I liked him.	**7.** Tôi thích ông ở điểm này.

trường hợp, có thể một câu tiếng Anh liên kết với hai hoặc nhiều câu tiếng Việt và ngược lại. Nói chung, có sáu loại quan hệ giữa các câu song ngữ [21], bao gồm:

1. *1:1* Các câu liên kết một-một.
2. *1:n* Một câu tiếng Anh liên kết với nhiều hơn một câu tiếng Việt.
3. *m:1* Nhiều hơn một câu tiếng Anh liên kết với một câu tiếng Việt.
4. *m:n* Nhiều hơn một câu tiếng Anh liên kết với nhiều hơn một câu tiếng Việt.
5. *m:0* Câu tiếng Anh không có câu tiếng Việt tương ứng.
6. *0:n* Câu tiếng Việt không có câu tiếng Anh tương ứng.

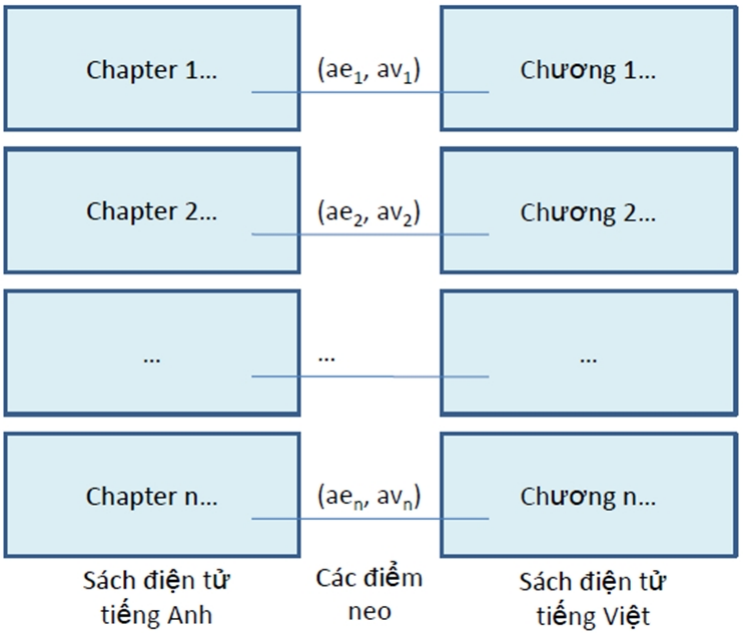
2.2.1 Tiền xử lý

Sách điện tử ban đầu ở định dạng PDF được chuyển đổi sang định dạng Text. Tuy nhiên, các thông tin về ranh giới đoạn bị mất trong quá trình chuyển đổi. Vì vậy chúng tôi phải phục hồi ranh giới đoạn. Để làm công việc này, trước tiên chúng tôi sử dụng một bộ công cụ có sẵn để chuyển đổi sách điện tử từ định dạng PDF sang Text. Sau đó, chúng tôi tiến hành khôi phục lại ranh giới giữa các đoạn. Bảng 2.4 là một ví dụ về phục hồi cấu trúc của đoạn văn bản gốc.

Bảng 2.4: Ví dụ minh họa ranh giới đoạn bị mất (trong quá trình chuyển đổi định dạng từ PDF sang Text) và được phục hồi.

Ranh giới đoạn bị mất
Your Highnesses, as Catholic Christians, and princes who love and promote the holy Christian faith, and are enemies of the doctrine of Mahomet, and of all idolatry and heresy, determined to send me, Christopher Columbus, to the above-mentioned countries of India, to see the said princes, people, and territories, and to learn their disposition and the proper method of converting them to our holy faith; and furthermore directed that I should not proceed by land to the East, as is customary, but by a Westerly route, in which direction we have hitherto no certain evidence that anyone has gone.
Ranh giới đoạn được phục hồi
Your Highnesses, as Catholic Christians, and princes who love and promote the holy Christian faith, and are enemies of the doctrine of Mahomet, and of all idolatry and heresy, determined to send me, Christopher Columbus, to the above-mentioned countries of India, to see the said princes, people, and territories, and to learn their disposition and the proper method of converting them to our holy faith; and furthermore directed that I should not proceed by land to the East, as is customary, but by a Westerly route, in which direction we have hitherto no certain evidence that anyone has gone.

Tiếp theo, chúng tôi sử dụng một hệ thống SMT để dịch văn bản trong sách tiếng Anh sang tiếng Việt. Quá trình này được sử dụng để đo độ tương tự giữa các đoạn/câu trong sách tiếng Anh và tiếng Việt. Bằng cách sử dụng SMT chúng tôi có thể làm giảm sự nhập nhằng về từ vựng trong bản dịch. Điều này thường xảy ra trong một số các nghiên cứu trước đây [37] khi sử dụng từ điển song ngữ.



Hình 2.3: Ví dụ về các điểm neo.

Trong phương pháp của chúng tôi, các điểm neo được sử dụng để hạn chế ảnh hưởng của lỗi hàng loạt (có thể xảy ra) trong quá trình gióng đoạn/câu. Với cách làm này, nếu có lỗi xảy ra, nó sẽ không ảnh hưởng đến gióng hàng trong toàn bộ cuốn sách. Thuật toán gióng đoạn/câu sẽ được thực hiện với các khối văn bản giữa hai điểm neo. Một số công trình trước [37, 105, 106] đo độ tương tự giữa bản dịch của văn bản nguồn và văn bản đích để xác định các điểm neo (ví dụ, bằng cách sử dụng điểm BLEU). Các điểm neo này là những khối văn bản ngắn với độ tương tự cao. Tuy nhiên, bằng cách này, chúng ta có thể không đạt được kết quả mong muốn. Ví dụ, vị trí của hai điểm neo liền kề quá xa hoặc quá gần. Trong một số trường hợp, các điểm neo được xác định không chính xác. Tất cả những vấn đề này có thể làm giảm độ chính xác của gióng hàng.

Theo quan sát của chúng tôi, trong hầu hết sách điện tử song ngữ Anh - Việt, một số khối văn bản có thể được sử dụng như điểm neo: *"Part"* $\leftrightarrow$ *"Phần"*, *"Chapter"* $\leftrightarrow$ *"Chương"*, vv. Chúng ta có thể thấy trong Hình 2.3, các điểm neo (ae_1, av_1), (ae_2, av_2), ..., (ae_n, av_n) được xác định bởi các khối văn bản: *"Chapter 1"* $\leftrightarrow$ *"Chương 1"*, *"Chapter 2"* $\leftrightarrow$ *"Chương 2"*, ..., *"Chapter n"* $\leftrightarrow$ *"Chương n"*. Ở đây, khối văn bản giữa hai điểm neo là toàn bộ một chương hoặc một phần trong cuốn sách. Lưu ý rằng, số lượng các điểm neo trong một cuốn sách điện tử là không nhiều. Do đó, chúng ta có thể dễ dàng phát hiện chúng bằng cách thủ công. Trong phương pháp của chúng tôi, các khối văn bản như mô tả ở trên có thể

được sử dụng để tự động phát hiện các điểm neo. Để làm công việc này, chúng tôi thực hiện hai bước sau:

1. Tạo ra danh sách L chứa các khối văn bản (xác định trước) như là các mẫu (pattern).
2. Với mỗi khối văn bản $l \in L$, nếu l khớp với một khối văn bản trong sách điện tử thì nó được xác định là điểm neo.

2.2.2 Đo độ tương tự

Giả sử chúng ta đang làm việc với sách điện tử song ngữ Anh - Việt. Sách tiếng Anh $\boldsymbol{E}$ chứa I khối (văn bản) $ue_1, ..., ue_I$ và sách tiếng Việt $\boldsymbol{V}$ chứa J khối $uv_1, ..., uv_J$. Gọi $\boldsymbol{T}$ là bản dịch tiếng Việt của $\boldsymbol{E}$ và ut_i là bản dịch tiếng Việt của khối ue_i (trong $\boldsymbol{E}$). Gọi $S_n(ut_i)$ và $D_n(uv_j)$ lần lượt là các tập $n-gram$ của các khối ut_i và uv_j. Độ tương tự giữa các khối ut_i và uv_j được định nghĩa như trong công thức (2.4).

$$Similarity_n(ut_i, uv_j) = \frac{|S_n(ut_i) \cap D_n(uv_j)|}{|S_n(ut_i) \cup D_n(uv_j)|} \tag{2.4}$$

Trong công thức này, $Similarity_n(ut_i, uv_j)$ là độ tương tự giữa hai khối văn bản ut_i và uv_j khi phân chia theo n, $0 \leq Similarity_n(ut_i, uv_j) \leq 1$.

2.2.3 Gióng hàng đoạn

Theo quan sát của chúng tôi, sự tương ứng của đoạn trong văn bản nguồn và đích không chỉ là $1-1$. Có năm loại phổ biến: $1-1$, $1-2$, $1-3$, $3-1$ và $2-1$. Các trường hợp khác thường không xảy ra. Dưới đây là một ví dụ về một đoạn văn bản tiếng Anh (trong cuốn sách Steve Jobs của tác giả Walter Isaacson) liên kết với hai đoạn trong bản dịch tiếng Việt:

- Văn bản tiếng Anh (trang 65).
  - *There was another reason that Joanne was balky about signing the adoption papers. Her father was about to die, and she planned to marry Jandali soon after. She held out hope, she would later tell family members,*

sometimes tearing up at the memory, that once they were married, she could get their baby boy back.

- Văn bản tiếng Việt (trang 21).
  - *Cũng còn một lý do khác khiến Joanne lúc đầu khăng khăng không ký giấy chuyển nhận con nuôi đó là vì cha bà sắp chết và bà dự định sẽ kết hôn với Jandali ngay sau đó.*
  - *Bà hi vọng rằng sau khi cưới nhau, họ sẽ thuyết phục dần được gia đình và nhận lại con.*

Chúng ta nhớ lại rằng, mục tiêu của giai đoạn này là tìm gióng hàng đoạn, liên kết các đoạn trong $\boldsymbol{E}$ với các đoạn trong $\boldsymbol{V}$. Với mỗi liên kết, chúng ta cần phát hiện ra các khối song ngữ (bao gồm một hoặc nhiều đoạn). Để thực hiện công việc này, chúng tôi tính toán độ tương tự của các khối theo các mẫu $1-1$, $1-2$, $1-3$, $2-1$ và $3-1$ bằng cách sử dụng hàm $Similarity_n(ut_i, uv_j)$ như trong công thức (2.4). Sau đó, cặp khối (u_s, u_t) có độ tương tự lớn nhất sẽ được chọn theo công thức (2.5).

$$(u_s, u_t) = \arg\max \begin{cases} Similarity_n(pt_i, pv_j) \\ Similarity_n(pt_i, pv_j pv_{j+1}) \\ Similarity_n(pt_i, pv_j pv_{j+1} pv_{j+2}) \\ Similarity_n(pt_i pt_{i+1}, pv_j) \\ Similarity_n(pt_i pt_{i+1} pt_{i+2}, pv_j) \end{cases} \tag{2.5}$$

2.2.4 Gióng hàng câu

Trong giai đoạn này, chúng ta xem xét làm thế nào để các câu có thể được liên kết trong một đoạn song ngữ (đã đạt được ở bước trước). Giả sử chúng ta có một đoạn song ngữ Anh - Việt (pe, pv). Trong đó, đoạn pe chứa k câu $se_1, ..., se_k$ và đoạn pv chứa m câu $sv_1, ..., sv_m$. Gọi $pt = st_1, ..., st_k$ là bản dịch tiếng Việt của đoạn pe. Nhiệm vụ của chúng ta trong giai đoạn này là tìm gióng hàng, liên kết các câu trong cặp đoạn (pe, pv).

Từ các khảo sát, chúng tôi thấy rằng mỗi câu trong đoạn nguồn chỉ có thể liên kết với các câu ở vị trí gần với nó trong đoạn đích. Thông thường, câu nguồn tại vị trí i (trong đoạn pe) thường liên kết với các câu tại các vị trí j, $(j+1)$, $(j+2)$

Thuật toán 2.1 Gióng hàng câu song ngữ cho sách điện tử.

- Đầu vào: $\boldsymbol{E}, \boldsymbol{V}, \boldsymbol{T}$
- Đầu ra: $\boldsymbol{A}$
  - Giai đoạn 1: liên kết các đoạn song ngữ
    1. Tìm vị trí của các điểm neo
    2. Với mỗi khối giữa hai điểm neo:
       (a) Tính toán độ tương tự của các khối theo các mẫu $1-1$, $1-2$, $1-3$, $2-1$ và $3-1$ dùng công thức (2.4).
       (b) Chọn cặp (u_s, u_t) có độ tương tự tốt nhất dùng công thức (2.5).
       (c) Liên kết u_s (trong $\boldsymbol{E}$) với u_t (trong $\boldsymbol{V}$).
  - Giai đoạn 2: gióng hàng câu cho các đoạn song ngữ
    1. $\boldsymbol{A} \leftarrow \emptyset$
    2. Với mỗi đoạn song ngữ (pe, pv) trong $(\boldsymbol{E}, \boldsymbol{V})$
       (a) Tách các đoạn pt, pv thành các câu (pt là bản dịch tiếng Việt của pe): $pt = st_1 st_2 \dots st_k$ và $pv = sv_1 sv_2 \dots sv_m$
       (b) Tính toán độ tương tự của các cặp câu (st_i, sv_j), (st_i, sv_{j+1}), (st_i, sv_{j+2}), (st_{i+1}, sv_j), (st_{i+2}, sv_j), $(st_i, sv_j sv_{j+1})$, $(st_i, sv_j sv_{j+1} sv_{j+2})$, $(st_i st_{i+1}, sv_j)$ và $(st_i st_{i+1} st_{i+2}, sv_j)$ dùng công thức (2.4).
       (c) Chọn cặp câu tại các vị trí (x, y) có độ tương tự tốt nhất dùng công thức (2.6).
       (d) Thêm liên kết $l = (x, y)$ vào $\boldsymbol{A}$: gióng hàng câu se_x với câu sv_y.

(trong đoạn pv) và ngược lại. Ở đây, $i = 1, \dots, k-2$ và $j = 1, \dots, m-2$. Các câu ở vị trí xa hơn cực kỳ hiếm và trên thực tế không xảy ra trong các thực nghiệm của chúng tôi. Ngoài ra, sự tương ứng của câu không chỉ là $1-1$. Tức là, có nhiều loại liên kết khác giữa các câu song ngữ, bao gồm: $1-2$, $1-3$, $3-1$ và $2-1$ (chúng ta có thể thấy trong Bảng 2.3). Do đó, chúng tôi chuyển đổi các liên kết này sang $1-1$ bằng cách ghép câu.

Tổng quát, nhiệm vụ của chúng ta là cần tìm ra câu ở vị trí thứ x ở trong đoạn pe là dịch của câu ở vị trí thứ y ở trong đoạn pv. Để làm điều này, chúng tôi tính toán độ tương tự của các cặp câu (st_i, sv_j), (st_i, sv_{j+1}), (st_i, sv_{j+2}), (st_{i+1}, sv_j), (st_{i+2}, sv_j), $(st_i, sv_j sv_{j+1})$, $(st_i, sv_j sv_{j+1} sv_{j+2})$, $(st_i st_{i+1}, sv_j)$ và $(st_i st_{i+1} st_{i+2}, sv_j)$. Sau đó, cặp câu (se_x, sv_y) có độ tương tự lớn nhất sẽ được lựa chọn như trong

công thức (2.6).

$$(se_x, sv_y) = \arg\max \begin{cases} Similarity_n(st_i, sv_j) \\ Similarity_n(st_i, sv_{j+1}) \\ Similarity_n(st_i, sv_{j+2}) \\ Similarity_n(st_{i+1}, sv_j) \\ Similarity_n(st_{i+2}, sv_j) \\ Similarity_n(st_i, sv_j sv_{j+1}) \\ Similarity_n(st_i, sv_j sv_{j+1} sv_{j+2}) \\ Similarity_n(st_i st_{i+1}, sv_j) \\ Similarity_n(st_i st_{i+1} st_{i+2}, sv_j) \end{cases} \tag{2.6}$$

Thuật toán 2.1 mô tả các bước để gióng hàng câu. Đầu vào cho thuật toán là một sách điện tử song ngữ Anh - Việt $(\boldsymbol{E}, \boldsymbol{V})$ và bản dịch tiếng Việt $\boldsymbol{T}$ của sách tiếng Anh $\boldsymbol{E}$. Đầu ra của thuật toán là gióng hàng $\boldsymbol{A}$ giữa các câu.

2.3 Thực nghiệm

2.3.1 Thực nghiệm về rút trích văn bản song ngữ từ Web

2.3.1.1 Cài đặt thực nghiệm

Để đánh giá hiệu quả của việc rút trích văn bản song ngữ từ Web, chúng tôi sử dụng các độ đo *precision* và *recall* như sau:

$$Precision = \frac{|X \cap Y|}{|X|} \tag{2.7}$$

$$Recall = \frac{|X \cap Y|}{|Y|} \tag{2.8}$$

Trong đó,

- X là tập hợp các cặp trang *web* được gán nhãn 1 bởi hệ thống (theo phương pháp được sử dụng).

- Y là tập hợp các cặp trang *web* được gán nhãn 1 trong tập dữ liệu kiểm tra (gán nhãn thủ công).

Ngoài ra, để cân bằng giữa độ chính xác và độ bao phủ, chúng sử dụng độ đo F_{score} như sau:

$$F_{score} = \frac{2 * Precision * Recall}{Precision + Recall} \tag{2.9}$$

Chúng tôi đã khảo sát các *web-site* tin tức song ngữ Anh - Việt trên World Wide Web. Một số *web-site* có chất lượng dịch tốt như BBC, VOA, VietnamPlus. Chúng tôi sử dụng công cụ Teleport-Pro để tải về 64.323 trang *web* (bao gồm cả tiếng Anh và tiếng Việt) từ ba *web-site* này theo các URL như trình bày ở Bảng 2.5. Trong đó 37.665 trang từ BBC, 12.553 trang từ VietnamPlus và 14.105 trang từ VOA News (chi tiết trong Bảng 2.6).

Bảng 2.5: Các URL từ ba *web-site*: BBC, VOA News và VietnamPlus.

Web-site	**URL tiếng Anh**	**URL tiếng Việt**
BBC	http://www.bbc.co.uk	http://www.bbc.co.uk/vietnamese/
VOA News	http://www.voanews.com	http://www.voanews.com/vietnamese/
VietnamPlus	http://en.vietnamplus.vn	http://www.vietnamplus.vn

Tiếp theo, chúng tôi tạo ra các cặp ứng viên từ nguồn dữ liệu thu thập được sử dụng một số ngưỡng: $sim_{cognate} > 0,5$ và $distance_{date} \leq 1$. Để xác định các giá trị ngưỡng này, chúng tôi tính các độ đo $sim_{cognate}$ và $distance_{date}$ trên tập dữ liệu gồm 50 trang *web* song ngữ Anh - Việt.

Bảng 2.6: Tổng hợp số trang *web* được tải về và số cặp ứng viên.

Web-site	**Số trang tải về**	**Số cặp ứng viên**
BBC	37.665	721
VOA News	14.105	129
VietnamPlus	12.553	320

Kết quả, chúng tôi đã loại bỏ hơn 90% cặp không được xem là ứng viên. Từ đó, chúng tôi nhận được 1.170 cặp ứng viên để xác định mỗi cặp trong số đó là song

```
<label> <index1>:<value1> <index2>:<value2>...
0    1:0.5205479452054795 2:0.5882352965397923 3:1 4:1
1    1:0.16470588235294115 2:0.5294117674740484 3:1 4:0
0    1:0.5862068965517242 2:0.09090909917355361 3:1 4:1
0    1:0.045454545454545414 2:0.37500000781249987 3:0 4:0
0    1:0.3695652173913043 2:0.3846153893491123 3:0 4:0
0    1:0.17948717948717952 2:0.20000000799999984 3:1 4:0
0    1:0.10948905109489049 2:0.4666666702222222 3:0 4:1
1    1:0.22651933701657456 2:0.1666666712962963 3:1 4:0
0    1:0.20346320346320346 2:0.5000000020833333 3:0 4:0
1    1:0.29166666666666663 2:0.7307692318047337 3:1 4:0
0    1:0.24255319148936172 2:0.4000000024 3:0 4:1
0    1:0.12195121951219512 2:0.5555555580246914 3:1 4:0
0    1:0.3513513513513513 2:0.6153846183431952 3:1 4:0
1    1:0.18213058419243988 2:0.5000000015625 3:1 4:0
0    1:0.49397590361445787 2:0.500000003125 3:1 4:1
1    1:0.23699421965317924 2:0.5333333364444444 3:0 4:0
1    1:0.125 2:0.2857142959183673 3:0 4:1
0    1:0.03448275862068961 2:0.2500000093749998 3:0 4:1
1    1:0.18959107806691455 2:0.42424242598714423 3:0 4:1
1    1:0.06976744186046513 2:0.3571428617346938 3:0 4:0
```

Hình 2.4: Định dạng dữ liệu huấn luyện phù hợp cho việc sử dụng công cụ LIBSVM.

ngữ hay không. Tiếp theo, chúng tôi thiết kế các đặc trưng về nội dung và cấu trúc cho tất cả các cặp ứng viên như trình bày ở các phần trước. Sau đó, chúng tôi thực hiện gán nhãn 0 hoặc 1 cho mỗi cặp ứng viên. Một cặp được gán nhãn bằng 1 nếu nó là song ngữ, ngược lại nó được gán nhãn 0. Có 433 cặp được gán nhãn 1 và 737 cặp có nhãn 0 từ 1.170 cặp ứng viên. Sau đó, chúng tôi xây dựng dữ liệu huấn luyện từ tập này với định dạng như trình bày ở Hình 2.4 - định dạng này phù hợp cho việc sử dụng công cụ LIBSVM[8]. Chúng tôi sử dụng kỹ thuật kiểm tra chéo 5 lần (5-folds cross-validation), mỗi phần (fold) có 234 cặp làm dữ liệu đánh giá và 936 cặp làm dữ liệu huấn luyện. Để đánh giá hiệu quả của phương pháp đã đề xuất, chúng tôi so sánh với hai cách tiếp cận trước đó: dựa trên cấu trúc [100] và dựa trên nội dung [76].

2.3.1.2 Kết quả thực nghiệm

Chúng tôi tiến hành các thực nghiệm với bốn phương pháp, cụ thể như sau:

[8]http://www.csie.ntu.edu.tw/~cjlin/libsvm/

1. Chỉ sử dụng các đặc trưng dựa trên cấu trúc theo phương pháp của Resnik [100] (hệ thống STRAND). Kết quả thực nghiệm được thể hiện ở Bảng 2.7.

Bảng 2.7: Kết quả thực nghiệm theo phương pháp của Resnik.

Phần	**Precision**	**Recall**	F_{Score}
Phần 1	40,9%	62,0%	49,3%
Phần 2	51,8%	61,4%	56,2%
Phần 3	39,7%	61,4%	48,2%
Phần 4	45,1%	76,3%	56,7%
Phần 5	44,4%	65,4%	52,9%
Trung bình	44,4%	65,3%	52,9%

2. Sử dụng từ điển song ngữ theo phương pháp của Ma [76] để so khớp các cặp từ (hệ thống BITS). Các kết quả thực nghiệm được trình bày ở Bảng 2.8.

Bảng 2.8: Kết quả thực nghiệm theo phương pháp của Ma.

Phần	**Precision**	**Recall**	F_{Score}
Phần 1	68,8%	48,4%	56,8%
Phần 2	64,7%	47,8%	55,0%
Phần 3	64,3%	54,8%	59,1%
Phần 4	60,1%	56,9%	58,4%
Phần 5	68,2%	52,8%	59,5%
Trung bình	65,2%	52,1%	57,8%

3. Kết hợp các đặc trưng dựa trên cấu trúc với các đặc trưng dựa vào nội dung sử dụng *cognate*. Các kết quả thực nghiệm được thể hiện ở Bảng 2.9.

Bảng 2.9: Kết quả thực nghiệm 3.

Phần	**Precision**	**Recall**	F_{Score}
Phần 1	87,3%	81,7%	84,4%
Phần 2	86,2%	84,2%	85,2%
Phần 3	86,9%	87,9%	87,4%
Phần 4	90,4%	81,7%	85,8%
Phần 5	90,4%	73,3%	81,0%
Trung bình	88,2%	81,7%	84,8%

4. Sử dụng các đặc trưng dựa trên cấu trúc với các đặc trưng dựa vào nội dung sử dụng phương pháp xác định phân đoạn dịch. Các kết quả thực nghiệm được thể hiện ở Bảng 2.10.

Bảng 2.10: Kết quả thực nghiệm 4.

Phần	**Precision**	**Recall**	F_{Score}
Phần 1	92,0%	81,6%	86,5%
Phần 2	91,1%	75,3%	82,4%
Phần 3	89,7%	77,7%	83,3%
Phần 4	90,4%	80,2%	85,0%
Phần 5	86,6%	77,5%	81,8%
Trung bình	90,0%	78,5%	83,8%

Điều đáng lưu ý là với bài toán này (rút trích văn bản song ngữ), độ chính xác là tiêu chí quan trọng nhất để đánh giá hiệu quả của một phương pháp. Các kết quả thực nghiệm cho thấy, hai phương pháp chúng tôi đề xuất đạt được kết quả tốt hơn (độ chính xác lần lượt là 88,2% và 90,0%) so với phương pháp sử dụng các đặc trưng dựa vào cấu trúc trang *web* của Resnik [100] (độ chính xác 44,4%) và phương pháp sử dụng từ điển của Ma [76] (độ chính xác 65,2%). Trong đó, kết hợp các đặc trưng dựa trên cấu trúc với các đặc trưng sử dụng *cognate* có độ chính xác là 88,2%; sử dụng các đặc trưng dựa trên cấu trúc với các đặc trưng theo phương pháp xác định phân đoạn dịch đạt được độ chính xác là 90,0%.

Theo các kết quả thực nghiệm, phương pháp đề xuất khá thành công trong việc rút trích các văn bản song ngữ từ Web. Những kết quả này đã cho thấy rằng các đặc trưng dựa trên nội dung như chúng tôi đưa ra là khá hiệu quả. Lưu ý rằng, nếu chúng ta không chắc chắn sự tương ứng về cấu trúc giữa hai trang *web*, chúng ta có thể chỉ sử dụng các đặc trưng dựa trên nội dung.

2.3.2 Thực nghiệm về rút trích câu song ngữ từ sách điện tử

2.3.2.1 Cài đặt thực nghiệm

Chúng tôi sử dụng bốn cuốn sách điện tử song ngữ Anh - Việt làm dữ liệu thực nghiệm. Các tiêu đề tiếng Anh của chúng lần lượt là: *Steve Jobs* (tác giả Walter

Isaacson), *The Open Society And Its Enemies* (tác giả Karl R. Popper), *The World is Flat* and *The Lexus and the Olive Tree* (tác giả Thomas L. Friedman). Các sách điện tử này được thu thập từ Internet và được dịch cẩn thận bởi các dịch giả nổi tiếng tại Việt Nam[9].

Bảng 2.11 trình bày thông tin chi tiết về sách điện tử song ngữ Anh - Việt được sử dụng trong thực nghiệm. Kích thước dữ liệu khoảng 10,1 MB (chỉ dữ liệu Text). Chúng bao gồm 5.381 đoạn văn bản tiếng Anh và 5.591 đoạn văn bản tiếng Việt. Trong bước tiền xử lý, chúng tôi chuyển đổi sách điện tử từ định dạng PDF sang Text. Để làm công việc này, đầu tiên chúng tôi sử dụng bộ công cụ *PDF to Text*[10] để chuyển đổi định dạng và sau đó phục hồi ranh giới đoạn. Trong bước tiếp theo, chúng tôi sử dụng *Google translator* như một hệ thống SMT để dịch văn bản (trong các cuốn sách) từ tiếng Anh sang tiếng Việt. Để đo độ tương tự giữa hai khối văn bản (ut_i and uv_j), chúng tôi sử dụng công thức (2.4) với $n = 1$.

Bảng 2.11: Thông tin chi tiết về sách điện tử song ngữ Anh - Việt được sử dụng trong thực nghiệm.

STT	**Tác giả**	**Tiếng Anh**	**Số đoạn**	**Tiếng Việt**	**Số đoạn**
1.	Walter Isaacson	Steve Jobs	1.968	Steve Jobs	1.948
2.	Karl R. Popper	The Open Society And Its Enemies	950	Xã Hội Mở Và Những Kẻ thù của Nó	904
3.	Thomas L. Friedman	The World is Flat	1.114	Thế giới Phẳng	1.348
4.	Thomas L. Friedman	The Lexus and the Olive Tree	1.349	Chiếc Lexus và Cây Ô Liu	1.391

[9] *Steve Jobs* được dịch bởi Bookstore Alezaa.com; *The Open Society And Its Enemies* được dịch bởi Nguyễn Quang A; *The World is Flat* và *The Lexus and the Olive Tree* được dịch bởi các dịch giả Nguyễn Quang A, Cao Việt Dung, Nguyễn Tiên Phong.

[10] http://www.pdf-technologies.com/pdf-library-pdf-to-text.aspx

2.3.2.2 Kết quả thực nghiệm

Chúng tôi chọn ngẫu nhiên 200 mẫu (của đoạn) từ dữ liệu thực nghiệm để đánh giá hiệu suất của phương pháp đã đề xuất. Kết quả thực nghiệm được trình bày trong Bảng 2.12. Độ chính xác đạt được là 97%. Kết quả thực nghiệm cho thấy rằng, phương pháp chúng tôi đề xuất là khá hiệu quả cho việc gióng hàng đoạn. Các đoạn song ngữ thu được chứa gần 40.000 câu song ngữ.

Bảng 2.12: Kết quả gióng hàng đoạn với 200 mẫu.

Mẫu	**Gióng hàng đúng**	**Gióng hàng sai**
1-1	158	2
1-2	16	2
1-3	0	0
2-1	16	2
3-1	4	0
Tổng	194	6

Như đã trình bày ở phần trước, sau khi gióng hàng đoạn, chúng tôi đã thu thập được ngữ liệu song ngữ Anh - Việt, trong đó có gần 40.000 câu gióng hàng ở mức đoạn. Để đánh giá tính hiệu quả của thuật toán gióng hàng câu, chúng tôi thiết kế bộ dữ liệu gồm 40 đoạn song ngữ từ bốn cuốn sách khác nhau như đã mô tả ở trên. Bảng 2.13 trình bày các kiểu quan hệ giữa các câu song ngữ của bộ dữ liệu này. Các kết quả thực nghiệm về gióng hàng câu được trình bày ở Bảng 2.14.

Bảng 2.13: Các kiểu quan hệ giữa các câu song ngữ trong 40 đoạn song ngữ.

Mẫu/Kiểu	**Số lượng**	**Tỷ lệ %**
1-1	175	86,6
1-2	11	5,4
1-3	2	1,0
2-1	13	6,4
3-1	1	0,5
Tổng	202	100

Cụ thể, phương pháp chúng tôi đề xuất đạt được độ chính xác 96,4%, độ bao phủ 93,6% và độ đo F_{score} 95,0%. Những kết quả này cho thấy, phương pháp chúng tôi đề xuất là khá hiệu quả trong việc gióng hàng câu cho sách điện tử song ngữ Anh - Việt. Bảng 2.15 cho thấy một số thông số của ngữ liệu đạt được từ phương pháp của chúng tôi.

Bảng 2.14: Kết quả thực nghiệm về gióng hàng câu.

Precision	**Recall**	F_{score}
96,4%	93,6%	95,0%

Bảng 2.15: Một số thống kê của ngữ liệu.

Tham số	**Tiếng Anh**	**Tiếng Việt**
Từ	771.565	1.035.358
Câu	39.066	36.104
Đoạn	5.042	5.042

2.3.3 Thực nghiệm về bổ sung ngữ liệu song ngữ cho dịch máy

Chúng tôi sử dụng ngữ liệu song ngữ Anh - Việt được tạo bởi Hoàng Cường và cộng sự [44]. Trong đó, tập dữ liệu huấn luyện gồm 90.000 câu song ngữ (kí hiệu là C_1) và tập dữ liệu gồm 1.000 câu song ngữ được sử dụng để đánh giá chất lượng dịch. Hệ thống SMT Anh - Việt dựa trên cụm từ được xây dựng với các thành phần như sau:

- Xây dựng mô hình ngôn ngữ với công cụ SRILM[11]. Chúng tôi xây dựng mô hình ngôn ngữ *3-gram* dùng 100.000 câu tiếng Việt.
- Xây dựng mô hình dịch và giải mã sử dụng công cụ MOSES[12] [61].

Trong thực nghiệm này, chúng tôi bổ sung 21.072 câu song ngữ Anh - Việt (kí hiệu là C_2) từ nguồn ngữ liệu song ngữ xây dựng được vào hệ thống dịch máy. Bảng 2.16 cho thấy các thông số về ngữ liệu song ngữ Anh - Việt được sử dụng trong thực nghiệm này. Trong tất cả các thực nghiệm về SMT thực hiện ở trong luận án này, chúng tôi sử dụng độ đo BLEU [93] để đánh giá chất lượng dịch.

Kết quả, chất lượng dịch của hệ thống SMT ban đầu (trên ngữ liệu C_1) đạt được 22,0 điểm BLEU. Sau khi chúng tôi bổ sung ngữ liệu C_2 với 21.072 câu song ngữ, chất lượng dịch tăng lên 3% (tương đương với 0,6 điểm BLEU).

[11]http://www.speech.sri.com/projects/srilm
[12]https://github.com/moses-smt/mosesdecoder

Bảng 2.16: Thống kê các thông số của ngữ liệu và chất lượng dịch của hệ thống.

Ngữ liệu	Tham số	Tiếng Anh	Tiếng Việt	Điểm BLEU
C_1	Số từ	1.136.973	1.165.361	22,0
(90.000)	Số từ vựng	46.033	44.050	
C_2	Số từ	398.776	41.0929	-
(21.072)	Số từ vựng	21.803	17.007	
$C_2 \cup C_2$	Số từ	1.535.749	1.576.290	22,6
(111.072)	Số từ vựng	54.884	51.713	

2.4 Kết luận chương

Chúng tôi đã trình bày các nội dung, kết quả nghiên cứu về xây dựng ngữ liệu song ngữ cho SMT. Trong nghiên cứu của chúng tôi, ngữ liệu song ngữ được khai thác từ Web và sách điện tử song ngữ. Từ thực nghiệm, chúng tôi thấy rằng, khai thác nguồn từ sách điện tử đạt được kết quả tốt hơn (độ chính xác cao hơn) so với nguồn từ Web. Nguyên nhân là nguồn từ sách điện tử được dịch cẩn thận (thường là dịch toàn bộ, sát nghĩa và ít lược bớt), trong khi đó nguồn từ Web thường bị nhiễu (có thể dịch toàn bộ hoặc chỉ dịch một số đoạn thậm chí chỉ dịch các thông tin chính). Các kết quả thực nghiệm cho thấy, chúng tôi có thể đạt được ngữ liệu song ngữ Anh - Việt đủ để xây dựng một hệ thống SMT thông qua việc khai thác ngữ liệu song ngữ từ hai nguồn này.

Đối với nguồn từ Web, chúng tôi đã kết hợp các đặc trưng dựa trên nội dung với các đặc trưng dựa trên cấu trúc của trang *web* để rút trích các văn bản song ngữ. Bài toán này được mô hình hóa như bài toán phân loại sử dụng phương pháp học máy dựa vào các đặc trưng giữa hai trang *web*. Chúng tôi đã đề xuất hai phương pháp thiết kế các đặc trưng dựa trên nội dung: (i) dựa trên *cognate* và (ii) dựa trên việc xác định các phân đoạn dịch. Đây là các phương pháp mới để đo độ tương tự về nội dung của hai trang *web* mà không đòi hỏi phải phân tích sâu về mặt ngôn ngữ. Kết quả thực nghiệm cho thấy các phương pháp đề xuất là khá thành công trong việc rút trích các văn bản song ngữ từ Web. Các kết quả thu được cũng cho thấy rằng, các đặc trưng dựa trên nội dung như đề xuất là thông tin quan trọng để xác định một cặp trang *web* là song ngữ hay không. Các phương pháp chúng tôi đã đề xuất có những ưu điểm sau:

- Thứ nhất, chúng tôi đã kết hợp cả đặc trưng về cấu trúc và đặc trưng về nội dung để tăng độ chính xác trong việc rút trích các văn bản song ngữ từ Web.
- Thứ hai, với việc sử dụng một hệ thống SMT ở phương pháp (ii), chúng ta có thể tận dụng những lợi thế của phương pháp dịch thống kê trong việc giải quyết các vấn đề về nhập nhằng từ vựng, dịch cụm từ và trật tự từ.
- Thứ ba, phương pháp (ii) có thể được áp dụng cho các cặp ngôn ngữ khác, vì rằng các đặc trưng được sử dụng trong phương pháp này là độc lập với ngôn ngữ.

Chúng tôi dự kiến sẽ tiếp tục công việc này với các thành phần song ngữ khác như đoạn, câu hoặc cụm từ. Công việc này cũng sẽ rất có ý nghĩa trong trường hợp chất lượng dịch giữa các trang *web* song ngữ không tốt. Ngoài ra, chúng tôi sẽ sử dụng hệ thống này để thu thập ngữ liệu song ngữ cho cặp ngôn ngữ Anh - Việt. Các kết quả thực nghiệm cho thấy rằng, có thể tự động xây dựng một kho ngữ liệu song ngữ Anh - Việt từ Web.

Đối với nguồn từ sách điện tử song ngữ, chúng tôi đã đề xuất một phương pháp hiệu quả để rút trích câu song ngữ (thông qua việc gióng hàng câu) cho cặp ngôn ngữ Anh - Việt. Để làm việc này, quá trình thực hiện trải qua hai giai đoạn: (i) liên kết các đoạn song ngữ và (ii) gióng hàng câu từ các đoạn song ngữ. Phương pháp của chúng tôi có ưu điểm là có thể phát hiện một số kiểu liên kết giữa các đoạn/câu song ngữ và giảm không gian tìm kiếm bằng cách sử dụng một số mẫu liên kết đoạn/câu. Bằng cách sử dụng phương pháp đã đề xuất, chúng tôi có thể đạt được ngữ liệu song ngữ gióng hàng ở mức câu đủ để xây dựng hệ thống SMT Anh - Việt.

Chương 3

Gióng hàng từ cho dịch máy thống kê

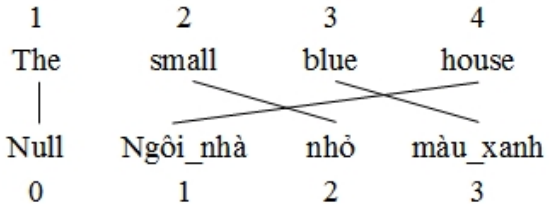
Chương này trình bày về gióng hàng từ cho SMT. Chúng tôi đề xuất một số cải tiến đối với mô hình IBM 1 theo cách tiếp cận dựa trên ràng buộc, bao gồm: ràng buộc neo, ràng buộc về vị trí của từ, ràng buộc về từ loại và ràng buộc về cụm từ. Với mỗi ràng buộc, chúng tôi đưa ra phương pháp tổng quát để tích hợp nó vào thuật toán EM trong quá trình ước lượng tham số của mô hình. Ngoài ra, chúng tôi đưa ra một phương pháp để kết hợp các ràng buộc. Các thực nghiệm, đánh giá về hiệu quả của gióng hàng từ cho SMT được trình bày ở cuối chương.

3.1 Cơ sở lý thuyết

3.1.1 Định nghĩa từ

Theo Diệp Quang Ban [6], về mặt nghiên cứu chung, người ta gặp không ít khó khăn trong việc xác định và nêu định nghĩa từ. Về mặt ngữ pháp, có thể hiểu từ là đơn vị nhỏ nhất có nghĩa và hoạt động tự do trong câu. Một số định nghĩa từ được hai tác giả Đinh Điền và Hồ Bảo Quốc trình bày trong [49] như sau:

- Theo L.Bloomfield, từ là "một hình thái tự do nhỏ nhất".
- Theo B.Golovin, từ là "đơn vị nhỏ nhất có nghĩa của ngôn ngữ, được vận dụng độc lập, tái hiện tự do trong lời nói để xây dựng nên câu".



HÌNH 3.1: Ví dụ về gióng hàng từ giữa một cặp câu song ngữ Anh - Việt.

- Theo Solncev, từ là "đơn vị ngôn ngữ có tính hai mặt: âm và nghĩa. Từ có khả năng độc lập về cú pháp khi sử dụng trong lời".

Đối với việc gióng hàng từ cho SMT, chúng ta cần xác định ranh giới từ. Ở đây, chúng tôi giả thiết đã có kết quả về phân đoạn từ (word segmentation) bằng cách sử dụng các công cụ phân đoạn từ từ các nghiên cứu trước [29, 47].

3.1.2 Định nghĩa bài toán gióng hàng từ

Cho câu **f** ở ngôn ngữ nguồn (câu nguồn) chứa J từ $f_1, ..., f_J$ và câu **e** ở ngôn ngữ đích (câu đích) chứa I từ $e_1, ..., e_I$, chúng tôi định nghĩa liên kết $l = (i, j)$ tồn tại nếu e_i và f_j là dịch (hoặc dịch một phần) của nhau. Khi đó, một gióng hàng từ **a** (giữa **f** và **e**) là một ánh xạ từ các vị trí từ trong **f** đến các vị trí từ trong **e** [60]:

$$\mathbf{a} : j \rightarrow i, \text{ với } j = 1, ..., J \text{ và } i = 0, ..., I \tag{3.1}$$

Trong gióng hàng từ **a**, mỗi a_j nhận một giá trị giữa 0 và I. Ở đây, J và I tương ứng là độ dài của câu nguồn và câu đích. Giá trị a_j biểu thị vị trí của từ đích e_{a_j} gióng hàng với từ nguồn f_j. Tức là, nếu một từ ở vị trí j trong câu nguồn được kết nối với một từ ở vị trí i trong câu đích thì $a_j = i$ và nếu không có kết nối đến bất kỳ từ nào ở câu đích thì $a_j = 0$ (nó sẽ được liên kết với từ *null*). Hình 3.1 minh họa gióng hàng từ giữa một cặp câu song ngữ Anh - Việt. Các cặp từ được gióng hàng bao gồm: (*house, ngôi_nhà*), (*small, nhỏ*), (*blue, màu_xanh*) và (*the, null*). Trong ví dụ này, gióng hàng từ **a** được biểu diễn như sau:

$$\mathbf{a} : \{1 \rightarrow 0, 2 \rightarrow 2, 3 \rightarrow 3, 4 \rightarrow 1\} \tag{3.2}$$

3.1.3 Các mô hình IBM

Brown và cộng sự [12] đề xuất một loạt năm mô hình thống kê (mô hình IBM 1-5) và cung cấp các thuật toán để ước lượng các tham số của những mô hình này. Các mô hình của Brown đã được sử dụng rộng rãi để gióng hàng từ cho SMT. Với mô hình IBM 1, xác suất $Pr(\mathbf{f}, \mathbf{a}|\mathbf{e})$ chỉ phụ thuộc vào tham số $t(f_j|e_i)$ - là xác suất từ f_j là bản dịch của từ e_{a_j}. Xác suất $Pr(\mathbf{f}, \mathbf{a}|\mathbf{e})$ được tính theo công thức (3.3)[1].

$$Pr(\mathbf{f}, \mathbf{a}|\mathbf{e}) = \frac{\varepsilon}{(I+1)^J} \prod_{j=1}^{J} t(f_j|e_{a_j}) \tag{3.3}$$

Mô hình IBM 2 sử dụng thêm tham số chuyển đổi trật tự từ cố định (absolute reordering) $a(i|j, I, J)$. Các mô hình IBM 3-5 bổ sung tham số độ dài của chuỗi từ được dịch $n(\phi|f)$ gọi là độ hỗn loạn, tức là số từ của ngôn ngữ đích mà từ của ngôn ngữ nguồn sinh ra. Tất cả các mô hình IBM đều sử dụng xác suất dịch từ vựng $t(f|e)$ từ mô hình IBM 1. Vì thế, việc tính toán xác suất này sẽ ảnh hưởng đến chất lượng gióng hàng từ đối với các mô hình IBM. Một trong những vấn đề đối với mô hình IBM 1 là không có các yếu tố để ngăn chặn các gióng hàng không mong muốn. Do đó, mỗi từ trong câu nguồn có thể gióng hàng với tất cả các từ ở trong câu đích. Điều này dẫn đến việc tính toán xác suất dịch từ vựng không chính xác. Để khắc phục hạn chế này, chúng tôi sử dụng một số ràng buộc để thu hẹp phạm vi gióng hàng. Tức là, mỗi từ trong câu nguồn chỉ gióng hàng với một (hoặc một số) từ trong câu đích nếu nó thỏa mãn ràng buộc nào đó. Chúng tôi sẽ trình bày chi tiết công việc này ở Phần 3.2.

3.1.4 Thuật toán cực đại kỳ vọng cho mô hình IBM 1

Thuật toán cực đại kỳ vọng [27] hay gọi tắt là thuật toán EM là một phương pháp ước lượng khả năng cực đại (Maximum Likelihood Estimation - MLE) hiệu quả trong bài toán dữ liệu ẩn. Trong MLE, chúng ta muốn ước lượng các tham số mô hình sao cho dữ liệu được quan sát là tương thích nhất. Mỗi vòng lặp của EM gồm hai bước: Bước E (expectation), dữ liệu ẩn được ước lượng dựa trên dữ liệu đã quan sát và các tham số của mô hình của ước lượng hiện tại. Bước M (maximization), hàm khả năng (likelihood function) được cực đại hóa với giả

[1] ε trong công thức (3.3) là hằng số chuẩn hóa, giá trị của ε được xác định bởi xác suất $p(J|I)$.

thuyết dữ liệu ẩn đã biết (sự ước lượng của dữ liệu ẩn trong bước E được sử dụng để thay thế dữ liệu ẩn thật sự). Thuật toán EM luôn hội tụ vì chắc chắn *likelihood* luôn tăng sau mỗi vòng lặp [86].

Ở đây, chúng tôi trình bày thuật toán EM cho mô hình IBM 1. Như đã giới thiệu ở Chương 1, xác suất $Pr(\mathbf{f}|\mathbf{e})$ được tính từ xác suất gióng hàng từ $Pr(\mathbf{f}, \mathbf{a}|\mathbf{e})$ theo công thức (1.3). Thuật toán này bao gồm hai bước sau:

- Bước E: Áp dụng mô hình trên dữ liệu, các xác suất gióng hàng được tính toán từ các tham số mô hình.

- Bước M: Ước lượng mô hình từ dữ liệu, giá trị của các tham số được ước lượng lại dựa trên các xác suất gióng hàng và dữ liệu.

Trong bước E, khi áp dụng mô hình trên dữ liệu, chúng ta cần tính xác suất của các gióng hàng khác nhau cho mỗi cặp câu trong dữ liệu. Tức là, chúng ta cần tính $Pr(\mathbf{a}|\mathbf{f}, \mathbf{e})$, xác suất của một gióng hàng cho cặp câu (**f**,**e**). Theo công thức Bayes, ta có:

$$Pr(\mathbf{a}|\mathbf{f}, \mathbf{e}) = \frac{Pr(\mathbf{f}, \mathbf{a}|\mathbf{e})}{Pr(\mathbf{f}|\mathbf{e})} \tag{3.4}$$

Chúng ta tính được $Pr(\mathbf{f}|\mathbf{e})$, xác suất dịch câu **e** sang câu **f** với gióng hàng bất kỳ như sau:

$$\begin{aligned}
Pr(\mathbf{f}|\mathbf{e}) &= \sum_{\mathbf{a}} Pr(\mathbf{f}, \mathbf{a}|\mathbf{e}) \\
&= \sum_{\mathbf{a}(1)=0}^{I} \cdots \sum_{\mathbf{a}(J)=0}^{I} Pr(\mathbf{f}, \mathbf{a}|\mathbf{e}) \\
&= \sum_{\mathbf{a}(1)=0}^{I} \cdots \sum_{\mathbf{a}(J)=0}^{I} \frac{\varepsilon}{(I+1)^J} \prod_{j=1}^{J} t(f_j|e_{a(j)}) \\
&= \frac{\varepsilon}{(I+1)^J} \sum_{\mathbf{a}(1)=0}^{I} \cdots \sum_{\mathbf{a}(J)=0}^{I} \prod_{j=1}^{J} t(f_j|e_{a(j)}) \\
&= \frac{\varepsilon}{(I+1)^J} \prod_{j=1}^{J} \sum_{i=0}^{I} t(f_j|e_i)
\end{aligned} \tag{3.5}$$

Lưu ý quan trọng ở bước biến đổi cuối cùng trong công thức (3.5). Thay vì thực hiện tính tổng trên I^J tích, chúng ta giảm việc tính toán xuống độ phức tạp tuyến

tính trong I và J. Kết hợp công thức (3.4) với (3.5), ta có:

$$\begin{aligned} Pr(\mathbf{a}|\mathbf{f},\mathbf{e}) &= \frac{Pr(\mathbf{f},\mathbf{a}|\mathbf{e})}{Pr(\mathbf{f}|\mathbf{e})} \\ &= \frac{\frac{\varepsilon}{(I+1)^J}\prod_{j=1}^{J} t(f_j|e_{a(j)})}{\frac{\varepsilon}{(I+1)^J}\prod_{j=1}^{J}\sum_{i=0}^{I} t(f_j|e_i)} \\ &= \prod_{j=1}^{J}\frac{t(f_j|e_{a(j)})}{\sum_{i=0}^{I} t(f_j|e_i)} \end{aligned} \tag{3.6}$$

Như vậy, chúng ta đã đặt cơ sở toán học cho bước E trong thuật toán EM. Công thức (3.6) định nghĩa làm thế nào để áp dụng mô hình trên dữ liệu.

Trong bước M, chúng ta cần thu thập số lượng dịch từ vựng (collect count) trên tất cả các gióng hàng có thể, giá trị này được xác định bởi xác suất của nó. Với mục đích này, chúng ta định nghĩa hàm c, hàm này thực hiện việc đếm số lần từ e dịch sang từ f ở trong cặp câu (**f**,**e**):

$$c(f|e;\mathbf{f},\mathbf{e}) = \sum_{\mathbf{a}} Pr(\mathbf{a}|\mathbf{f},\mathbf{e})\sum_{j=1}^{J}\delta(f,f_j)\delta(e,e_{a(j)}) \tag{3.7}$$

Trong công thức (3.7), hàm Kronecker $\delta(x,y)$ nhận giá trị là 1 nếu x = y và 0 nếu ngược lại.

Thay giá trị $Pr(\mathbf{a}|\mathbf{f},\mathbf{e})$ từ công thức (3.6) vào công thức (3.7) và thực hiện tối giản tương tự như trong công thức (3.6), ta có:

$$c(f|e;\mathbf{f},\mathbf{e}) = \frac{t(f|e)}{\sum_{i=0}^{I} t(f|e_i)}\sum_{j=1}^{J}\delta(f,f_j)\sum_{i=0}^{I}\delta(e,e_i) \tag{3.8}$$

Bây giờ, chúng ta có thể ước lượng phân phối xác suất dịch mới theo công thức (3.9).

$$t(f|e;\mathbf{f},\mathbf{e}) = \frac{\sum_{(\mathbf{f},\mathbf{e})} c(f|e;\mathbf{f},\mathbf{e})}{\sum_{e}\sum_{(\mathbf{f},\mathbf{e})} c(f|e;\mathbf{f},\mathbf{e})} \tag{3.9}$$

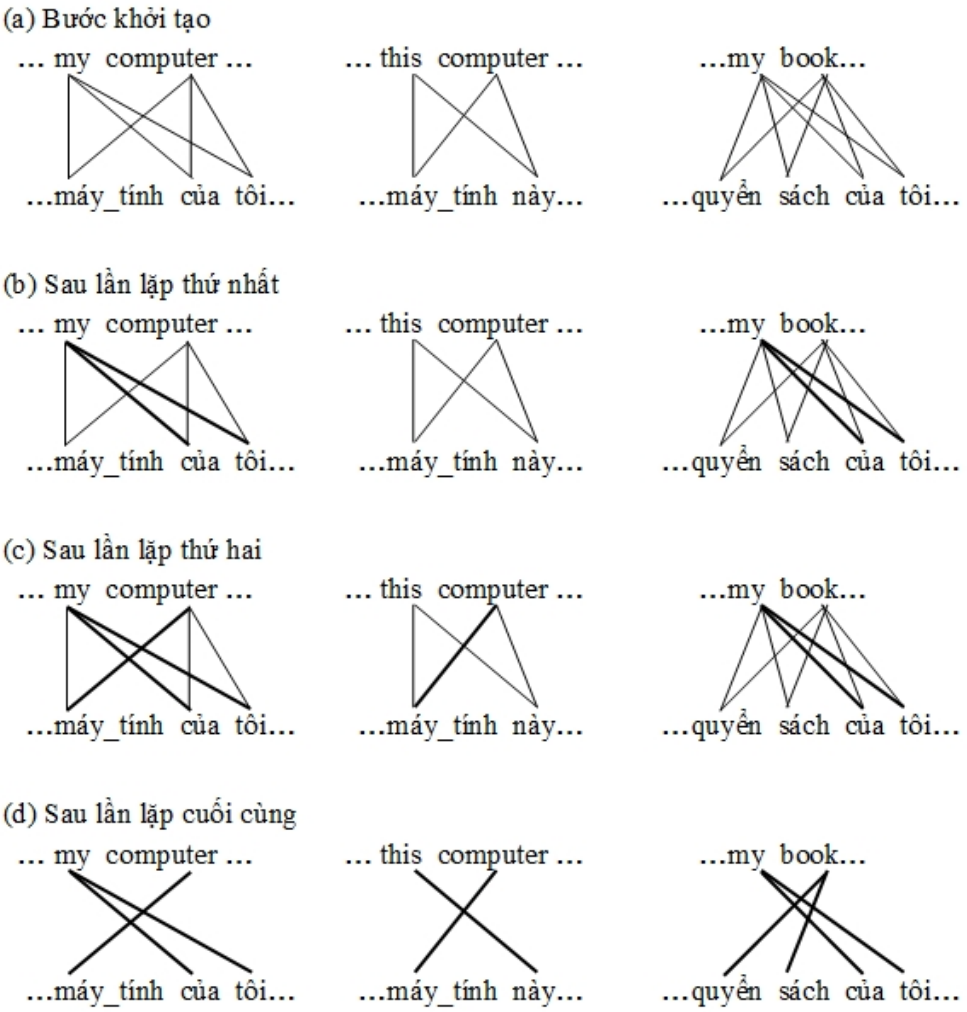
Thuật toán 3.1 trình bày cài đặt giả mã (pseudo-code) của thuật toán EM cho mô hình IBM 1.

Thuật toán 3.1 Thuật toán EM cho mô hình IBM 1 [60].

Đầu vào: Ngữ liệu C={(**f**,**e**)}
Đầu ra: Xác suất dịch $t(f|e)$

```
khởi tạo t(f|e) (phân phối đều)
while not (hội tụ) do
    //khởi tạo
    count(f|e) = 0 for all f, e
    total(e) = 0 for all e
    for all các cặp câu (f,e) do
        //compute normalization
        for all từ f in f do
            s_total(f) = 0
            for all từ e in e do
                s_total(f) += t(f|e)
            end for
        end for
        //đếm số lần từ e dịch sang từ f
        for all từ f in f do
            for all từ e in e do
                count(f|e) += t(f|e)/s_total(f)
                total(e) += t(f|e)/s_total(f)
            end for
        end for
    end for
    //ước lượng xác suất t(f|e)
    for all từ f ở ngôn ngữ nguồn do
        for all từ e ở ngôn ngữ đích do
            t(f|e) = count(f|e)/total(e)
        end for
    end for
end while
```

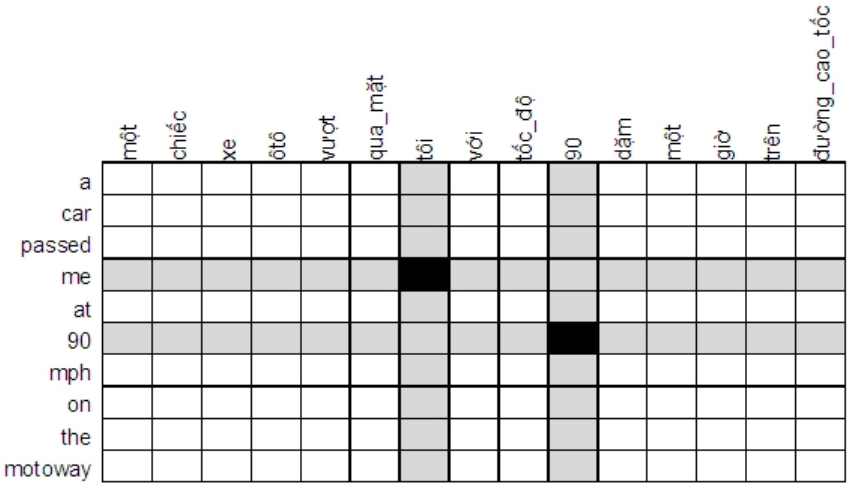
Ví dụ với cặp ngôn ngữ Anh - Việt, giả sử ta có tập dữ liệu huấn luyện C = {(*...my computer...*, *...máy_tính của tôi...*), (*...this computer...* ,*...máy_tính này...*), (*...my book...*, *...quyển sách của tôi...*)}. Hình 3.2 minh họa quá trình gióng hàng từ trên tập dữ liệu huấn luyện C theo thuật toán EM. Ở bước khởi tạo (hình (a)), mỗi từ ở câu nguồn đều có khả năng gióng hàng đến tất cả các từ ở câu đích. Sau lần lặp đầu tiên (hình (b)), liên kết từ "*my*" và "*của tôi*" được xác định. Ở lần lặp kế tiếp: liên kết từ "*computer*" và "*máy tính*" như hình (c). Thêm một lần lặp nữa, liên kết từ "*this*" và "*này*", "*book*" và "*quyển sách*" dựa theo nguyên lý "chuồng bồ câu" (pigeon hole principle). Cuối cùng ta có kết quả gióng hàng từ như hình (d).



Hình 3.2: Minh họa quá trình gióng hàng từ theo thuật toán EM.

3.2 Một số cải tiến mô hình IBM 1 theo cách tiếp cận dựa trên ràng buộc

Trong phần này, chúng tôi trình bày một số cải tiến đối với mô hình IBM 1 theo cách tiếp cận dựa trên ràng buộc. Ngoại trừ ràng buộc neo đã được đề cập trong nghiên cứu của Talbot [111], chúng tôi đề xuất ba ràng buộc mới, đó là ràng buộc về vị trí của từ, ràng buộc về từ loại và ràng buộc về cụm từ. Với mỗi ràng buộc, chúng tôi đưa ra phương pháp tổng quát để tích hợp nó vào thuật toán EM trong quá trình ước lượng tham số của mô hình. Ngoài ra, chúng tôi đưa ra một phương pháp để kết hợp các ràng buộc. Với việc sử dụng ràng buộc neo, nghiên cứu của chúng tôi khác với Talbot [111] ở hai điểm: (i) chúng tôi xác định các điểm neo hoàn toàn dựa vào dữ liệu huấn luyện thay vì sử dụng thêm từ điển song ngữ;



Hình 3.3: Ví dụ về ràng buộc neo (ô màu đen), gán xác suất gióng hàng bằng không cho tất cả các cặp từ khác (ô màu xám).

(ii) chúng tôi đưa ra phương pháp tổng quát để tích hợp ràng buộc này vào thuật toán EM.

3.2.1 Cải tiến mô hình IBM 1 sử dụng ràng buộc neo

Ràng buộc neo (anchor constraint) là ràng buộc loại trừ, trong đó nó tạo ra một gióng hàng tin cậy giữa hai từ [111]. Gióng hàng giữa hai từ tại một điểm neo được tạo ra bằng cách thiết lập xác suất gióng hàng bằng không ở vị trí đó cho tất cả các từ khác. Hình 3.3 cho thấy một ví dụ về ràng buộc neo. Như chúng ta thấy trong hình, cặp từ *(tôi, me)* là một điểm neo, do đó xác suất gióng hàng giữa *tôi* và các từ khác như *(tôi, a), (tôi, car), (tôi, passed)*,... được gán bằng không.

Để tạo ra các điểm neo, chúng tôi sử dụng *cognate* (như trình bày ở Phần 2.1.2.1). Thông tin này đặc biệt hữu ích khi từ điển song ngữ điện tử không có sẵn. Chúng tôi khác với phương pháp của Kondrak trong [64] - tác giả đã sử dụng ba độ đo về sự tương tự giữa các từ: Simard, hệ số Dice và LCSR để xác định *cognate*. Talbot [111] sử dụng từ điển song ngữ để tạo ra các điểm neo. Ở đây, chúng tôi lựa chọn những từ không được dịch và nó cùng xuất hiện trong cặp câu song ngữ (ví dụ: chữ viết tắt, chữ số,...). Lưu ý rằng trong phương pháp của chúng tôi, *cognate* được xác định từ dữ liệu huấn luyện.

Ngoài việc sử dụng *cognate* làm điểm neo, chúng tôi sử dụng thêm các cặp từ song ngữ (từ dữ liệu huấn luyện). Để xác định các cặp từ này, chúng tôi kết hợp

giữa xác suất dịch từ vựng ($t(f_j|e_i)$) và tần suất xuất hiện ($count(f_j, e_i)$) của nó trong dữ liệu huấn luyện. Chúng tôi định nghĩa danh sách L là tập hợp các cặp từ song ngữ như sau:

$$L = \{(f_j, e_i)|t(f_j|e_i) > \alpha, count(f_j, e_i) > \beta\}. \tag{3.10}$$

Trong đó, e_i là từ ở ngôn ngữ nguồn, f_j là từ ở ngôn ngữ đích và α, β là các ngưỡng được xác định trước.

Bây giờ, chúng tôi tạo ra ràng buộc dựa trên các điểm neo và biểu diễn nó bởi hàm $anchor\_constraint(f_j, e_i)$, như sau:

$$anchor\_constraint(f_j, e_i) = \begin{cases} true & \text{nếu } (f_j = e_i) \vee (f_j, e_i) \in L \\ false & \text{ngược lại} \end{cases} \tag{3.11}$$

Chúng tôi tích hợp ràng buộc neo vào bước M trong thuật toán EM bằng cách định nghĩa lại hàm c (ở công thức (3.8)). Cụ thể, nếu cặp từ (f, e) trong cặp câu (**f**,**e**) là một điểm neo, chúng tôi gán xác suất gióng hàng bằng 1 cho cặp từ này (hay nói cách khác, thiết lập xác suất gióng hàng bằng không ở vị trí đó cho tất cả các từ khác). Trong trường hợp này, hàm c nhận giá trị là 1.

Tổng quát, gọi C_{anchor} là ràng buộc neo. Với mỗi cặp từ (f, e) trong cặp câu (**f**,**e**), chúng tôi tính giá trị của hàm c theo hai trường hợp như sau:

$$c(f|e; \mathbf{f}, \mathbf{e}, C_{anchor}) = \begin{cases} 1 \text{ nếu } anchor\_constraint(f, e) = true \\ \frac{t(f|e)}{\sum_{i=0}^{I} t(f|e_i)} \sum_{j=1}^{J} \delta(f, f_j) \sum_{i=0}^{I} \delta(e, e_i) \qquad \text{ngược lại} \end{cases} \tag{3.12}$$

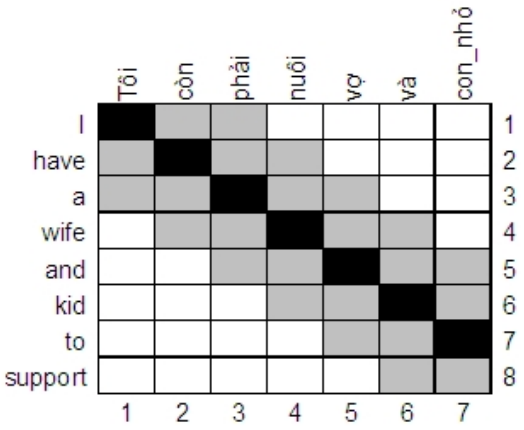
Sau khi tính giá trị của hàm $c(f|e; \mathbf{f}, \mathbf{e}, C_{anchor})$ trên toàn bộ tập dữ liệu huấn luyện, chúng tôi ước lượng xác suất dịch $t(f|e)$ theo công thức (3.9)). Thuật toán 3.2 trình bày quá trình tích hợp ràng buộc neo vào thuật toán EM cho mô hình IBM 1.

Thuật toán 3.2 Thuật toán EM cho mô hình IBM 1 sử dụng ràng buộc neo.

Đầu vào: Ngữ liệu C={(**f**,**e**)}, danh sách L
Đầu ra: Xác suất dịch $t(f|e)$

1: khởi tạo $t(f|e)$ (phân phối đều)
2: **while** not (hội tụ) **do**

```
    //khởi tạo
    count(f|e) = 0 for all f, e
    total(e) = 0 for all e
    for all các cặp câu (f,e) do
      //compute normalization
      anchor_points[j] = false for all j = 1, .., J
      for all từ f_j in f do
        s_total(f_j) = 0
        for all từ e_i in e do
          s_total(f_j) += t(f_j|e_i)
          if anchor_constraint(f_j, e_i) then
            anchor_points[j] = true
            s_total(f_j) = t(f_j|e_i)
            break
          end if
        end for
      end for
      //đếm số lần từ e dịch sang từ f
      for all từ f_j in f do
        if anchor_points[j] then
          for all từ e_i in e do
            if anchor_constraint(f_j, e_i) then
              count(f_j|e_i) += t(f_j|e_i)/s_total(f_j)
              total(e_i) += t(f_j|e_i)/s_total(f_j)
              break
            end if
          end for
        else
          for all từ e_i in e do
            count(f_j|e_i) += t(f_j|e_i))/s_total(f_j)
            total(e_i) += t(f_j|e_i))/s_total(f_j)
          end for
        end if
      end for
    end for
    //ước lượng xác suất t(f|e)
    for all từ f ở ngôn ngữ nguồn do
      for all từ e ở ngôn ngữ đích do
        t(f|e) = count(f|e)/total(e)
      end for
    end for
  end while
```



Hình 3.4: Ví dụ về ràng buộc về vị trí của từ với ngưỡng $\delta = 2$, mỗi vị trí đích j (ô màu đen) chỉ gióng hàng với các vị trí nguồn ở trong phạm vi $[j - \delta, j + \delta]$ (ô màu xám).

3.2.2 Cải tiến mô hình IBM 1 sử dụng ràng buộc về vị trí của từ

Từ thực tế chúng tôi thấy rằng, các từ trong câu nguồn (câu tiếng Anh) thường gióng hàng với các từ trong câu đích (câu tiếng Việt) ở vị trí gần nó. Theo quan sát này, chúng tôi đề xuất một ràng buộc mới dựa vào khoảng cách giữa vị trí của từ trong câu nguồn và từ trong câu đích ở trong câu song ngữ. Chúng tôi biểu diễn ràng buộc này bởi hàm $distance\_constraint(i, j)$, như sau:

$$distance\_constraint(i, j) = \begin{cases} true & \text{nếu } abs(i - j) \leq \delta \\ false & \text{ngược lại} \end{cases} \tag{3.13}$$

Trong đó, $abs(i - j)$ là khoảng cách từ vị trí nguồn i đến vị trí đích j và δ là ngưỡng được xác định trước. Điều này có nghĩa rằng, cho một cặp câu $(\mathbf{f}, \mathbf{e})$, mỗi vị trí đích chỉ gióng hàng với các vị trí nguồn ở trong phạm vi $[j - \delta, j + \delta]$. Chúng ta có thể thấy trong Hình 3.4, từ đích ở vị trí 5 (từ *vợ*) chỉ gióng hàng với những từ nguồn tại các vị trí trong phạm vi $[3, 7]$ (các từ *a, wife, and, kid, to*).

Tương tự như với ràng buộc neo, ràng buộc về vị trí của từ được chúng tôi tích hợp vào bước M trong thuật toán EM bằng cách định nghĩa lại hàm c (ở công thức (3.8)). Với mỗi cặp từ (f, e) trong cặp câu $(\mathbf{f},\mathbf{e})$, chúng tôi gán trọng số cao hơn nếu ràng buộc về vị trí của từ được thỏa mãn và trọng số thấp hơn trong trường hợp ngược lại. Tức là, xác suất gióng hàng giữa f và e được nhân với trọng

số λ khi ràng buộc được thỏa mãn và nhân với $(1-\lambda)$ nếu ràng buộc không thỏa mãn.

Tổng quát, chúng tôi gọi $C_{distance}$ là ràng buộc về vị trí của từ áp dụng lên mỗi cặp từ (f, e) trong cặp câu (**f**,**e**). Khi đó, hàm c được định nghĩa lại như sau:

$$c(f|e;\mathbf{f},\mathbf{e},C_{distance}) = \begin{cases} \lambda \frac{t(f|e)}{\sum_{i=0}^{I} t(f|e_i)} \sum_{j=1}^{J} \delta(f,f_j) \sum_{i=0}^{I} \delta(e,e_i) & \text{nếu } abs(i-j) \leq \delta \\ (1-\lambda) \frac{t(f|e)}{\sum_{i=0}^{I} t(f|e_i)} \sum_{j=1}^{J} \delta(f,f_j) \sum_{i=0}^{I} \delta(e,e_i) & \text{ngược lại} \end{cases} \tag{3.14}$$

Sau khi tính giá trị của $c(f|e;\mathbf{f},\mathbf{e},C_{distance})$ trên toàn bộ tập dữ liệu huấn luyện, chúng tôi ước lượng xác suất dịch $t(f|e)$ theo công thức (3.9)). Thuật toán 3.3 trình bày quá trình tích hợp ràng buộc về vị trí của từ vào thuật toán EM cho mô hình IBM 1.

Thuật toán 3.3 Thuật toán EM cho mô hình IBM 1 sử dụng ràng buộc về từ vị trí của từ.

Đầu vào: Ngữ liệu C={(**f**,**e**)}
Đầu ra: Xác suất dịch $t(f|e)$

khởi tạo $t(f|e)$ (phân phối đều), $lamda = 0,99$
while not (hội tụ) **do**
  //khởi tạo
  $count(f|e) = 0$ for all f, e
  $total(e) = 0$ for all e
  for all các cặp câu (**f**,**e**) **do**
    //compute normalization
    for all từ f in **f** **do**
      $s\_total(f) = 0$
      for all từ e in **e** **do**
        $s\_total(f) += t(f|e)$
      end for
    end for
    //đếm số lần từ e dịch sang từ f
    for all từ f_j in **f** **do**
      for all từ e_i in **e** **do**
        if $dis\_constraint(i,j)$ **then**
          $count(f|e) += (lamda * t(f|e))/s\_total(f)$
          $total(e) += (lamda * t(f|e))/s\_total(f)$
        else
          $count(f|e) += ((1-lamda) * t(f|e))/s\_total(f)$
          $total(e) += ((1-lamda) * t(f|e))/s\_total(f)$
        end if
      end for
    end for
  end for

```
    //ước lượng xác suất t(f|e)
    for all từ f ở ngôn ngữ nguồn do
        for all từ e ở ngôn ngữ đích do
            t(f|e) = count(f|e)/total(e)
        end for
    end for
end while
```

3.2.3 Cải tiến mô hình IBM 1 sử dụng ràng buộc về từ loại

3.2.3.1 Quan hệ về từ loại

Theo cách hiểu thông thường, trong một câu song ngữ, các từ có cùng từ loại thường gióng hàng với nhau hơn so với các từ khác nhau về từ loại. Tương tự Lee [67], trong nghiên cứu này, chúng tôi giả thuyết rằng tất cả các nhãn từ loại (Part Of Speech - POS) ở ngôn ngữ nguồn có một số quan hệ với các POS ở ngôn ngữ đích.

Ký hiệu R là tập hợp các quan hệ về POS giữa tiếng Anh và tiếng Việt, như sau:

$$R = \{(x \rightarrow y) | x \in X, y \in Y\} \tag{3.15}$$

Trong đó, X và Y tương ứng là tập chứa các thẻ POS của tiếng Anh và tiếng Việt. Chúng tôi xây dựng tập hợp các quan hệ POS giữa tiếng Anh và tiếng Việt. Để có được mối quan hệ POS, chúng tôi ước lượng xác suất $Pr(y|x)$ từ dữ liệu huấn luyện và chọn ra các cặp thẻ (x, y) sao cho xác suất $Pr(y|x) > \theta$, với θ là ngưỡng được xác định trước. Bảng 3.1 cho thấy một số quan hệ về POS giữa tiếng Anh và tiếng Việt.

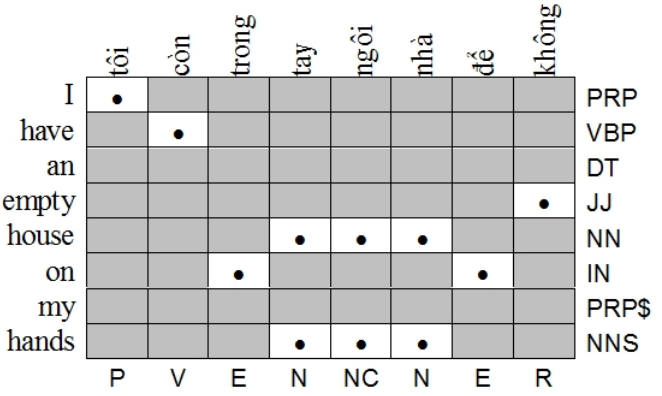
3.2.3.2 Ràng buộc về từ loại

Ràng buộc POS đòi hỏi mỗi từ nguồn f_j chỉ gióng hàng với các từ đích e_i có cùng quan hệ về POS. Chúng ta có thể thấy trong Hình 3.5, các cặp từ thỏa mãn ràng buộc POS, như dưới đây:

- i/PRP → tôi/P

Bảng 3.1: Một số quan hệ về POS giữa tiếng Anh và tiếng Việt theo xác suất.

STT	x	y	Pr(y|x)
1.	CC	C	0,10312350218036845
2.	CC	CC	0,64422509506983890
3.	CD	M	0,42324043405388434
4.	CD	Nu	0,34604169428201010
5.	DT	L	0,07183175269876728
6.	DT	Null	0,09736918511791175
7.	IN	E	0,24146922356101302
8.	IN	C	0,10751210932995296
9.	JJR	A	0,01617469492564175
10.	JJS	A	0,01554019785021995
11.	JJ	A	0,20368923073348824
12.	NN	N	0,32408358694587674
13.	PRP	P	0,30907509774229660
14.	VBD	V	0,05963677111170705
15.	VBP	V	0,01427621927014257



Hình 3.5: Ví dụ về ràng buộc từ loại (chấm tròn đen), gán xác suất dịch bằng 0 cho tất cả các cặp từ khác (ô màu xám).

- have/VBP → còn/V
- an/DT → *null*
- empty/JJ → không/R
- house/NN → tay/N, ngôi/Nc, nhà/N
- on/IN → trong/E, để/E
- my/PRP$ → *null*

- hands/NNS → tay/N, ngôi/Nc, nhà/N

Ký hiệu $P(f_j)$, $P(e_i)$ tương ứng với thẻ POS của từ nguồn f_j và từ đích e_i. Khi đó, một cặp từ (f_j, e_i) thỏa mãn ràng buộc POS nếu $P(f_j) \rightarrow P(e_i) \in R$. Chúng tôi biểu diễn ràng buộc POS bởi hàm $pos\_constraint(f_j, e_i)$, như sau:

$$pos\_constraint(f_j, e_i) = \begin{cases} true & \text{nếu } P(f_j) \rightarrow P(e_i) \in R \\ false & \text{ngược lại} \end{cases} \tag{3.16}$$

Tương tự như với ràng buộc neo và ràng buộc về vị trí của từ, chúng tôi tích hợp ràng buộc từ loại vào bước M trong thuật toán EM bằng cách định nghĩa lại hàm c (ở công thức (3.8)). Với mỗi cặp từ (f, e) trong cặp câu (**f**,**e**), gọi $E_{pos} = \{e_1, e_2, ..., e_n\}$ là tập hợp các từ trong **e** thỏa mãn ràng buộc từ loại. Điều này có nghĩa là $P(f) \rightarrow P(e_k) \in R, 1 \leq k \leq n$.

Tổng quát, nếu cặp từ (f, e) trong cặp câu (**f**,**e**) thỏa mãn ràng buộc C_{pos} (ở đây là ràng buộc về từ loại) thì hàm c được định nghĩa như sau:

$$c(f|e; \mathbf{f}, \mathbf{e}, C_{pos}) = \frac{t(f|e)}{\sum_{e_k \in E_{pos}} t(f|e_k)} \sum_{j=1}^{J} \delta(f, f_j) \sum_{i=0}^{I} \delta(e, e_i) \tag{3.17}$$

Chi tiết quá trình tích hợp ràng buộc từ loại vào thuật toán EM cho mô hình IBM 1 được thể hiện ở Thuật toán 3.4. Chúng ta có thể thấy trong thuật toán, hàm $pos\_contraint(f, e)$ (ở các dòng 11 và 19) sẽ hạn chế mỗi từ nguồn f chỉ gióng hàng với các từ đích e chỉ khi ràng buộc POS được thỏa mãn.

Thuật toán 3.4 Thuật toán EM cho mô hình IBM 1 sử dụng ràng buộc về từ loại.

Đầu vào: Ngữ liệu C={(**f**,**e**)} đã được gán nhãn từ loại
Đầu ra: Xác suất dịch $t(f|e)$

1: khởi tạo $t(f|e)$ (phân phối đều)
2: **while** not (hội tụ) **do**
3: //khởi tạo
4: $count(f|e) = 0$ for all f, e
5: $total(e) = 0$ for all e

```
    for all các cặp câu (f,e) do
      //compute normalization
      for all từ f in f do
        s_total(f) = 0
        for all từ e in e do
          if pos_constraint(f, e) then
            s_total(f)+ = t(f|e)
          end if
        end for
      end for
      //đếm số lần từ e dịch sang từ f
      for all từ f in f do
        for all từ e in e do
          if pos_constraint(f, e) then
            count(f|e)+ = t(f|e)/s_total(f)
            total(e)+ = t(f|e)/s_total(f)
          end if
        end for
      end for
    end for
    //ước lượng xác suất t(f|e)
    for all từ f ở ngôn ngữ nguồn do
      for all từ e ở ngôn ngữ đích do
        t(f|e) = count(f|e)/total(e)
      end for
    end for
end while
```

3.2.4 Cải tiến mô hình IBM 1 sử dụng ràng buộc về cụm từ

Khi áp dụng các mô hình IBM để gióng hàng từ cho hệ thống SMT từ tiếng Anh sang tiếng Việt, chúng tôi thấy nhiều gióng hàng sai. Nhiều trong số đó là sự khác biệt nhiều về cú pháp. Quan sát này thúc đẩy chúng tôi sử dụng ràng buộc về cụm từ sử dụng các mẫu cú pháp song ngữ. Ràng buộc này yêu các từ trong cụm song ngữ gióng hàng với nhau và không gióng hàng với các từ khác bên ngoài cụm. Phương pháp của chúng tôi đưa ra bao gồm ba bước như sau: (i) định nghĩa tập các mẫu cú pháp song ngữ (bilingual syntactic pattern), (ii) sử dụng mẫu cú pháp song ngữ để xác định ràng buộc về cụm từ và (iii) tích hợp ràng buộc về cụm từ vào mô hình IBM 1.

3.2.4.1 Mẫu cú pháp song ngữ

Ở đây, chúng tôi chỉ xem xét các mẫu đơn giản và thường xuất hiện trong ngữ liệu. Ví dụ, một số mẫu cú pháp song ngữ là cụm danh từ tiếng Anh và tiếng Việt được trình bày như sau:

- DT(a, an) NN / M(một) Nc N
- DT(a, an) JJ* NN / M(một) Nc N A*
- DT(a, an) JJ* NN / M(một) N A*
- DT JJ* NNS / L N A*
- DT(this, that) NN / Nc N P
- DT(these, those) NNS / L N P

Lưu ý rằng mỗi mẫu bao gồm hai phần, cách nhau bởi một dấu gạch ngang. Phần bên trái mô tả mẫu cú pháp trong ngôn ngữ nguồn (tiếng Anh), phần bên phải mô tả mẫu cú pháp tương ứng trong ngôn ngữ đích (tiếng Việt). Trong hình thức thể hiện này, chúng tôi cũng sử dụng ký hiệu ({*}) để biểu diễn sự lặp lại của thẻ POS như trong ví dụ trên.

3.2.4.2 Ràng buộc về cụm từ

Giả sử rằng, chúng ta có cặp câu (**f**,**e**) trong ngữ liệu song ngữ so khớp với mẫu cú pháp song ngữ tại vị trí (j_1, j_2) ở câu nguồn và (i_1, i_2) ở câu đích. Bây giờ, chúng tôi tách mỗi câu thành ba phần $\mathbf{f} = \overline{f_1}, \overline{f_2}, \overline{f_3}$ và $\mathbf{e} = \overline{e_1}, \overline{e_2}, \overline{e_3}$. Lưu ý rằng, phần bên trái và bên phải có thể rỗng. Như vậy, ($\overline{f_2}$, $\overline{e_2}$) là cụm từ song ngữ. Ở đây, ràng buộc về cụm từ yêu cầu mỗi từ f_j trong cụm từ nguồn $\overline{f_2}$ chỉ gióng hàng với các từ e_i trong cụm từ đích $\overline{e_2}$. Tương tự, các từ ngoài cụm từ nguồn gióng hàng với các từ ngoài cụm từ đích.

Tương tự như với các ràng buộc trước, chúng tôi tích hợp ràng buộc về cụm từ vào bước M trong thuật toán EM, chúng tôi định nghĩa lại hàm c (ở công thức (3.8)). Với mỗi từ f ở trong câu nguồn **f**, nếu $f \in \overline{f_2}$ thì hàm c được định nghĩa

Bảng 3.2: 13 mẫu cú pháp song ngữ Anh - Việt được sử dụng trong ràng buộc về cụm từ.

STT	Tiếng Anh	Tiếng Việt
1.	DT(a, an) NN	M(một) Nc N
2.	DT(this, that, these, those) NN	Nc N P
3.	DT NNS	L Nc N P
4.	DT(these, those) NNS	L N P
5.	DT(this, that) JJ NN	Nc N A P
6.	DT(a,an) JJ NN	M(một) Nc N A
7.	DT(a,an) JJ NN	M(một) N A
8.	DT JJ NNS	L N A
9.	PRP$ NN	Nc N E P
10.	RBR JJ	A R
11.	RBS JJ	A R
12.	PRP$ NNS	L N E P
13.	PRP$ JJ NN	N A E P

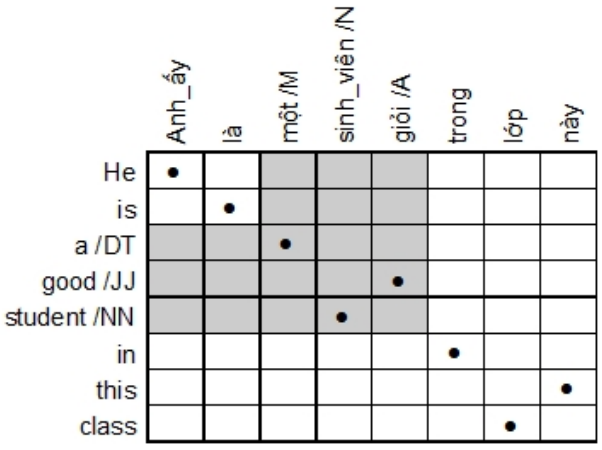
bởi công thức (3.18). Ngược lại, hàm c được định nghĩa theo công thức (3.19).

$$c(f|e;\overline{f_2},\overline{e_2},C_{phrase}) = \frac{t(f|e)}{\sum_{i=i_1}^{i_2} t(f|e_i)} \sum_{j=j_1}^{j_2} \delta(f,f_j) \sum_{i=i_1}^{i_2} \delta(e,e_i) \tag{3.18}$$

$$c(f|e;\overline{f_1 f_3},\overline{e_1 e_3},C_{phrase}) = \frac{t(f|e)}{\sum_{i\notin(i_1..i_2)} t(f|e_i)} \sum_{j\notin(j_1..j_2)} \delta(f,f_j) \sum_{i\notin(i_1..i_2)} \delta(e,e_i) \tag{3.19}$$

Kết hợp các công thức (3.18), (3.19) và (3.9), bây giờ chúng tôi ước lượng tham số $t(f|e)$ sử dụng ràng buộc về cụm từ như sau:

1. Khởi tạo $t(f|e)$ (phân phối đều).



HÌNH 3.6: Ví dụ về gióng hàng từ giữa một cặp câu Anh - Việt (các chấm tròn đen), các từ tiếng Anh và tiếng Việt được liệt kê tương ứng theo chiều dọc và chiều ngang. Các ô màu xám thể hiện ràng buộc về cụm từ.

2. Với mỗi cặp câu $(\mathbf{f}^{(s)}, \mathbf{e}^{(s)}), 1 \leq s \leq S$[2].

   - Với mỗi mẫu cú pháp trong tập các mẫu cú pháp đã định nghĩa trước.
     - Nếu mẫu cú pháp so khớp thì sử dụng các công thức (3.18) và (3.19) để tính giá trị của hàm c.
     - Ngược lại, hàm c được tính theo công thức (3.8).

3. Ước lượng giá trị mới cho tham số $t(f|e)$:

   - Với mỗi e ở trong $\mathbf{e}^{(s)}$,
   - Với mỗi f ở trong $\mathbf{f}^{(s)}$, sử dụng công thức (3.9) để đạt được giá trị mới cho $t(f|e)$.

4. Lặp lại các bước 2 và 3 cho đến khi giá trị của $t(f|e)$ hội tụ.

Chúng ta có thể thấy ở Hình 3.6, mẫu cú pháp song ngữ *DT(a) JJ NN / M(một) N A* so khớp ở trong cặp câu (*He is a good student in this class*, *Anh_ấy là một sinh_viên tốt trong lớp này*). Khi đó, câu nguồn được tách thành $\overline{f_1}$ = *"He is"*, $\overline{f_2}$ = *"a good student"* và $\overline{f_3}$ = *"in this class"*. Tương tự, câu đích phân chia thành $\overline{e_1}$ = *"Anh_ấy là"*, $\overline{e_2}$ = *"một sinh_viên tốt"* và $\overline{e_3}$ = *"trong lớp này"*.

[2] S là kích thước của dữ liệu huấn luyện.

3.2.5 Kết hợp các ràng buộc

Ở các phần trước, chúng tôi đã đề xuất một số ràng buộc và đưa ra cách để tích hợp chúng vào quá trình ước lượng tham số của mô hình IBM 1 trong thuật toán EM. Vấn đề đặt ra là làm thế nào để kết hợp chúng lại với nhau? Ở đây chúng tôi đưa ra một phương pháp để kết hợp các ràng buộc này, dùng "phép hợp" giữa các ràng buộc.

Gọi $C = \{c_1, c_2, ..., c_K\}$ là tập các ràng buộc. Cặp từ (f, e) (trong cặp câu (**f**,**e**)) được gọi là thỏa mãn ràng buộc nếu nó thỏa mãn một ràng buộc $c_k \in C$ bất kỳ, $1 \leq k \leq K$ (tức là, thỏa mãn ràng buộc c_1 hoặc c_2, ..., hoặc c_K). Như vậy, chúng tôi hạn chế mỗi từ f trong câu nguồn chỉ gióng hàng với các từ e trong câu đích nếu cặp từ (f, e) thỏa mãn ràng buộc. Gọi $E_C = \{e_1, e_2, ..., e_n\}$ là tập hợp các từ trong **e** thỏa mãn ràng buộc. Khi đó, hàm c được định nghĩa lại như sau:

$$c(f|e; \mathbf{f}, \mathbf{e}, C) = \frac{t(f|e)}{\sum_{e_k \in E_C} t(f|e_k)} \sum_{j=1}^{J} \delta(f, f_j) \sum_{i=0}^{I} \delta(e, e_i) \tag{3.20}$$

Về cơ bản, việc ước lượng xác suất dịch $t(f|e)$ và tích hợp tập ràng buộc C vào thuật toán EM cho mô hình IBM 1 được thực hiện tương tự như với các ràng buộc chúng tôi đã trình bày ở trên.

3.3 Thực nghiệm

3.3.1 Cài đặt thực nghiệm

Quá trình thực nghiệm, đánh giá về gióng hàng từ được thực hiện trên hệ thống SMT Anh - Việt (dịch từ tiếng Anh sang tiếng Việt). Để đánh giá hiệu quả của những cải tiến chúng tôi đã đề xuất, các thực nghiệm được thực hiện với ba loại: (i) mô hình IBM gốc, (ii) phương pháp gióng hàng từ không giám sát tốt nhất hiện nay (state of the art) [40, 74, 119] - Giza++ [91] và (iii) mô hình IBM cải tiến theo những đề xuất của chúng tôi. Ngoài ra, chúng tôi so sánh các kết quả thực nghiệm theo các phương pháp chúng tôi đã đề xuất với kết quả thực nghiệm của một số nghiên cứu gần đây về gióng hàng từ cho SMT (chi tiết chúng tôi sẽ trình bày trong Phần 3.3.5).

Chúng tôi sử dụng ngữ liệu song ngữ Anh - Việt được tạo bởi Hoàng Cường và cộng sự [44]. Chúng tôi thiết kế bốn tập dữ liệu huấn luyện lần lượt chứa 60.000, 70.000, 80.000 và 90.000 câu song ngữ Anh - Việt. Bảng 3.3 thống kê chi tiết về các thông số của dữ liệu này. Tập dữ liệu gồm 1.000 câu song ngữ Anh - Việt được sử dụng để đánh giá chất lượng dịch. Để gán nhãn từ loại cho dữ liệu huấn luyện, chúng tôi sử dụng công cụ *vnTagger* [3] cho văn bản tiếng Việt và *posTagger-1.0*[4] cho văn bản tiếng Anh. Chúng tôi cài đặt lại thuật toán EM[5] bằng cách tích hợp các ràng buộc như đã trình bày ở các phần trước vào quá trình ước lượng tham số của mô hình IBM 1. Với ràng buộc về cụm từ, chúng tôi xây dựng tập hợp các

Bảng 3.3: Thống kê ngữ liệu song ngữ Anh - Việt được sử dụng để xây dựng mô hình dịch.

Số cặp câu	**Tiếng Anh**		**Tiếng Việt**	
	Số từ	**Số từ vựng**	**Số từ**	**Số từ vựng**
60.000	762.725	37.458	774.572	34.981
70.000	888.999	40.197	906.467	37.626
80.000	1.013.492	44.062	1.037.375	41.888
90.000	1.136.973	46.033	1.165.361	44.050

mẫu cú pháp song ngữ Anh - Việt chứa 13 cặp mẫu như trình bày ở Bảng 3.4.

Hệ thống SMT dựa trên cụm từ được xây dựng với các thành phần như sau:

- Xây dựng mô hình ngôn ngữ với công cụ SRILM[6]. Chúng tôi xây dựng mô hình ngôn ngữ *3-gram* dùng 100.000 câu tiếng Việt.
- Xây dựng mô hình dịch và giải mã sử dụng công cụ MOSES[7] [61].

Trong tất cả các thực nghiệm dưới đây, chúng tôi thực hiện cùng một lược đồ huấn luyện với số lần theo trình tự như sau: 5 lần lặp mô hình IBM 1, 3 lần lặp mô hình IBM 2 và 3 lần lặp mô hình IBM 3. Trong các bảng từ 3.5 đến 3.9, các ký hiệu Δ_1 và Δ_2 lần lượt là độ chênh lệnh điểm BLEU (tăng (+)/giảm (−)) giữa phương pháp của chúng tôi so với phương pháp sử dụng mô hình IBM gốc và phương pháp sử dụng Giza++. Chúng tôi thực hiện kiểm chứng thống kê các kết quả

[3]http://mim.hus.vnu.edu.vn/phuonglh/softwares/vnTagger
[4]http://www-tsujii.is.s.u-tokyo.ac.jp/ tsuruoka/postagger/
[5]Các mô hình IBM gốc được cài đặt bởi Hoàng Cường và cộng sự trong [45].
[6]http://www.speech.sri.com/projects/srilm
[7]https://github.com/moses-smt/mosesdecoder

Bảng 3.4: Thống kê số lần đồng xuất hiện của 13 mẫu cú pháp song ngữ Anh- Việt.

STT	**Tiếng Anh**	**Tiếng Việt**	**Số lần đồng xuất hiện**
1.	DT(a, an)/NN	M(một)/Nc/N	1.600
2.	DT(this, that, these, those)/NN	Nc/N/P	701
3.	DT/NNS	L/Nc/N/P	67
4.	these, those/NNS	L/N/P	418
5.	DT(this, that)/JJ/NN	Nc/N/A/P	40
6.	DT(a,an)/JJ/NN	M(một)/Nc/N/A	321
7.	DT(a,an)/JJ/NN	M(một)/N/A	1.877
8.	DT/JJ/NNS	L/N/A	506
9.	PR P$/NN	Nc/N/E/P	133
10.	RBR/JJ	A/R	204
11.	RBS/JJ	A/R	102
12.	PRP$/NNS	L/N/E/P	208
13.	PRP$/JJ/NN	N/A/E/P	201

đạt được theo phương pháp ngẫu nhiên gần đúng (approximate randomization) sử dụng công cụ MultEval [19]. Độ tin cậy của các kết quả này thể hiện qua trị số p (hay còn gọi là *p-value*) trong các trường hợp phương pháp của chúng tôi đạt được điểm BLEU cao hơn mô hình IBM gốc hoặc Giza++ (tức là: $\Delta_1 > 0$ hoặc $\Delta_2 > 0$).

3.3.2 Kết quả thực nghiệm với ràng buộc neo và ràng buộc về vị trí của từ

Chúng tôi sử dụng các giá trị $\alpha = 0,5$, $\beta = 10$ với ràng buộc neo; $\delta = 2$ và $\lambda = 0,99$ với ràng buộc về vị trí. Bảng 3.5 và 3.6 trình bày kết quả thực nghiệm với các kích thước dữ liệu huấn luyện khác nhau. Chúng ta có thể thấy, mô hình IBM được cải tiến với việc sử dụng hai ràng buộc này đã đạt được điểm BLEU cao hơn so với mô hình IBM gốc trên cả bốn tập dữ liệu huấn luyện. Cụ thể, điểm BLEU tăng trung bình 0,67 điểm với ràng buộc neo (tương đương với việc chất lượng MT tăng 3,03%) và 1,48 điểm với ràng buộc về vị trí của từ (tương đương với việc chất lượng MT tăng 6,49%). Ngoài ra, so với Giza++, tính trung bình trên cả bốn tập dữ liệu, phương pháp của chúng tôi đạt được điểm BLEU cao hơn 0,28 điểm khi sử dụng ràng buộc neo và 1,08 điểm khi sử dụng ràng buộc về vị trí của từ. Đối với ràng buộc neo, khi kích thước dữ liệu huấn luyện tăng thì sự chênh lệch giữa phương pháp của chúng tôi so với Giza++ là không nhiều.

Bảng 3.5: Kết quả thực nghiệm với mô hình IBM gốc, Giza++ và sử dụng ràng buộc neo.

Kích thước ngữ liệu	Mô hình IBM gốc	Giza++	Ràng buộc neo	Δ_1 (trị số p)	Δ_2 (trị số p)
60.000	21,6	21,9	22,7	$+1,1_{(0,0001)}$	$+0,8_{(0,0264)}$
70.000	22,3	22,7	23,2	$+0,9_{(0,0016)}$	$+0,5_{(0,2160)}$
80.000	23,2	23,8	23,7	$+0,5_{(0,0434)}$	-0,1
90.000	23,6	23,9	23,8	$+0,2_{(0,1560)}$	-0,1

Bảng 3.6: Kết quả thực nghiệm với mô hình IBM gốc, Giza++ và sử dụng ràng buộc về vị trí của từ.

Kích thước ngữ liệu	Mô hình IBM gốc	Giza++	Ràng buộc về vị trí	Δ_1 (trị số p)	Δ_2 (trị số p)
60.000	21,6	21,9	22,9	$+1,3_{(0,0016)}$	$+1,0_{(0,0110)}$
70.000	22,3	22,7	23,7	$+1,4_{(0,0001)}$	$+1,0_{(0,0026)}$
80.000	23,2	23,8	24,8	$+1,6_{(0,0001)}$	$+1,0_{(0,0108)}$
90.000	23,6	23,9	25,2	$+1,6_{(0,0002)}$	$+1,3_{(0,0005)}$

3.3.3 Kết quả thực nghiệm với ràng buộc từ loại

Chúng tôi sử dụng ngưỡng $\theta = 0,01$ để xác định các quan hệ về từ loại. Bảng 3.7 trình bày kết quả thực nghiệm với mô hình IBM gốc, Giza++ và sử dụng ràng buộc về từ loại (mô hình IBM cải tiến) trên các tập ngữ liệu khác nhau. Kết quả thực nghiệm cho thấy, sử dụng ràng buộc về từ loại đạt được điểm BLEU cao hơn trên tất cả các tập dữ liệu huấn luyện so với mô hình IBM gốc và Giza++. Cụ thể, khi sử dụng ràng buộc về từ loại điểm BLEU tăng trung bình 0,98 điểm, tương đương với việc chất lượng MT tăng 4,31% so với mô hình IBM gốc. Ngoài ra, so với sử dụng Giza++, phương pháp dùng ràng buộc từ loại đạt được chất lượng dịch tốt hơn 2,50% (tương đương 0,58 điểm BLEU).

BẢNG 3.7: Kết quả thực nghiệm với mô hình IBM gốc, Giza++ và sử dụng ràng buộc về từ loại.

Kích thước ngữ liệu	Mô hình IBM gốc	Giza++	Ràng buộc về từ loại	Δ_1 (trị số p)	Δ_2 (trị số p)
60.000	21,6	21,9	22,4	$+0,8_{(0,0116)}$	$+0,5_{(0,1996)}$
70.000	22,3	22,7	23,6	$+1,3_{(0,0001)}$	$+0,9_{(0,0170)}$
80.000	23,2	23,8	24,1	$+0,9_{(0,0329)}$	$+0,3_{(0,5060)}$
90.000	23,6	23,9	24,5	$+0,9_{(0,0400)}$	$+0,6_{(0,1419)}$

3.3.4 Kết quả thực nghiệm với ràng buộc cụm từ

Bảng 3.8 trình bày kết quả thực nghiệm trên các tập dữ liệu huấn luyện khác nhau. Kết quả thực nghiệm cho thấy, cải tiến của chúng tôi đạt được điểm BLEU cao hơn so với mô hình IBM gốc trên tất cả các tập dữ liệu huấn luyện. Cụ thể, điểm BLEU tăng trung bình $0,45$ điểm so với mô hình IBM gốc không sử dụng ràng buộc. So sánh với Giza++, phương pháp dùng ràng buộc cụm từ đạt được điểm BLEU cao hơn trung bình $0,05$ điểm. Chúng tôi muốn nhấn mạnh rằng, trong các thực nghiệm với ràng buộc về cụm từ, chúng tôi sử dụng 13 mẫu cú pháp song ngữ. Chúng tôi tin tưởng rằng, phương pháp chúng tôi đưa ra có thể đạt được các kết quả tốt hơn nếu chúng ta tăng số lượng mẫu cú pháp song ngữ.

Bảng 3.8: Kết quả thực nghiệm với mô hình IBM gốc, Giza++ và sử dụng ràng buộc về cụm từ.

Kích thước ngữ liệu	Mô hình IBM gốc	Giza++	Ràng buộc về cụm từ	Δ_1 (trị số p)	Δ_2 (trị số p)
60.000	21,6	21,9	22,1	$+0{,}5_{(0{,}0164)}$	$+0{,}2_{(0{,}6501)}$
70.000	22,3	22,7	22,8	$+0{,}5_{(0{,}0140)}$	$+0{,}1_{(0{,}7222)}$
80.000	23,2	23,8	23,8	$+0{,}6_{(0{,}0115)}$	0,0
90.000	23,6	23,9	23,8	$+0{,}2_{(0{,}3883)}$	-0,1

3.3.5 Kết quả thực nghiệm về kết hợp ràng buộc

Trong thực nghiệm này, chúng tôi kết hợp giữa ràng buộc về vị trí của từ với ràng buộc về từ loại. Việc kết hợp hai ràng buộc này được thực hiện theo phương pháp như chúng tôi đã trình bày ở Phần 3.2.5. Bảng 3.9 cho thấy kết quả thực nghiệm về kết hợp ràng buộc. Chúng ta có thể thấy, mô hình IBM được cải tiến khi kết hợp ràng buộc đạt được điểm BLEU cao hơn mô hình IBM gốc và Giza++ trên tất cả các tập dữ liệu huấn luyện. Khi chúng tôi kết hợp ràng buộc về vị trí của từ với ràng buộc về từ loại, chất lượng dịch tốt hơn so với việc sử dụng riêng lẻ từng ràng buộc. Cụ thể, so với mô hình IBM gốc điểm BLEU tăng trung bình 1,63 điểm khi kết hợp ràng buộc, tương đương với việc chất lượng MT tăng 7,16% với độ tin cậy $p \leq 0,0007$. So với việc sử dụng Giza++, phương pháp kết hợp ràng buộc này đạt được điểm BLEU cao hơn trung bình 1,23 điểm với độ tin cậy $p \leq 0,0034$.

Bảng 3.9: Kết quả thực nghiệm với mô hình IBM gốc, Giza++ và kết hợp ràng buộc (vị trí của từ với từ loại).

Kích thước ngữ liệu	Mô hình IBM gốc	Giza++	Kết hợp ràng buộc	Δ_1 (trị số p)	Δ_2 (trị số p)
60.000	21,6	21,9	23,1	$+1{,}5_{(0{,}0007)}$	$+1{,}2_{(0{,}0034)}$
70.000	22,3	22,7	23,8	$+1{,}5_{(0{,}0002)}$	$+1{,}1_{(0{,}0020)}$
80.000	23,2	23,8	25,0	$+1{,}8_{(0{,}0001)}$	$+1{,}2_{(0{,}0019)}$
90.000	23,6	23,9	25,3	$+1{,}7_{(0{,}0002)}$	$+1{,}4_{(0{,}0004)}$

Như vậy, từ các kết quả thực nghiệm với các ràng buộc chúng tôi đề xuất, chúng ta có thể thấy chất lượng dịch tăng trung bình từ 0,45 đến 1,63 điểm BLEU so với mô hình IBM gốc và từ 0,05 đến 1,12 so với Giza++. Trong đó, sự kết hợp giữa các ràng buộc (ở đây là sự kết hợp giữa ràng buộc về vị trí của từ với ràng

buộc về từ loại) đạt được kết quả tốt hơn so với việc sử dụng từng ràng buộc riêng lẻ.

Ngoài ra, chúng tôi so sánh kết quả thực nghiệm theo phương pháp kết hợp ràng buộc với kết quả thực nghiệm của một số nghiên cứu gần đây về gióng hàng từ cho SMT:

1. Phương pháp gióng hàng từ của Songyot và Chiang trong [110] sử dụng thông tin học mô hình tương tự từ, từ dữ liệu đơn ngữ dựa trên mạng nơ-ron. Thông tin này sau đó được tích hợp vào các mô hình IBM.

2. Phương pháp gióng hàng từ không giám sát với các đặc trưng tùy ý được đề xuất bởi Chris Dyer và cộng sự trong [33].

Do sự khác biệt về phương pháp tiếp cận giữa hai nghiên cứu [110] và [33] với nghiên cứu của chúng tôi. Hơn nữa, mỗi phương pháp được cài đặt thử nghiệm trên các cặp ngôn ngữ khác nhau và nó không được chia sẻ rộng rãi. Cho nên, ở đây chúng tôi so sánh gián tiếp dựa trên các kết quả thực nghiệm được trình bày trong [110] và [33] với kết quả thực nghiệm của chúng tôi. Cơ sở của so sánh này là các kết quả thực nghiệm trong cả ba nghiên cứu đều được so sánh với cùng một *baseline* là Giza++. Bảng 3.10 cho thấy kết quả thực nghiệm của ba phương pháp (phương pháp của chúng tôi với hai phương pháp [110] và [33]). Ký hiệu Δ trong Bảng 3.10 là độ chênh lệnh điểm BLEU (tăng (+)/giảm (−)) giữa mỗi phương pháp gióng hàng so với Giza++. Trong đó, phương pháp của chúng tôi và phương pháp [110] có thực hiện kiểm chứng thống kê thông qua trị số p.

Chúng ta có thể thấy trên Bảng 3.10, phương pháp chúng tôi đề xuất về kết hợp ràng buộc trên cặp ngôn ngữ Anh - Việt đạt được kết quả tốt hơn hoặc bằng với các phương pháp [110] và [33]. Cụ thể, điểm BLEU của phương pháp chúng tôi tăng trung bình 1,2 điểm với độ tin cậy $p < 0,05$ và bằng phương pháp [110] trong trường hợp tốt nhất trên cặp ngôn ngữ Trung - Anh. Trong các trường hợp còn lại, phương pháp chúng tôi có điểm BLEU tăng cao hơn hai phương pháp [110] và [33] từ 0,1 đến 0,4 điểm.

Bảng 3.10: So sánh với một số nghiên cứu gần đây về gióng hàng từ cho SMT.

Phương pháp của Chris Dyer và cộng sự [33]				
Cặp ngôn ngữ	**Giza++**	Chris Dyer và cộng sự	Δ	**Tỷ lệ%**
Séc (Czech) - Anh	16,3	17,4	+1,1	6,75%
Urdu - Anh	23,3	24,1	+0,8	3,43%
Phương pháp của Songyot và Chiang [110]				
Cặp ngôn ngữ	**Giza++**	Songyot và Chiang	Δ (độ tin cậy)	**Tỷ lệ%**
Trung - Anh	22,0	23,2	$+1{,}2_{(p<0{,}05)}$	5,47%
Ả Rập - Anh	33,6	34,4	$+0{,}8_{(p<0{,}05)}$	2,53%
Phương pháp của chúng tôi				
Cặp ngôn ngữ	**Giza++**	**Kết hợp ràng buộc**	Δ (độ tin cậy)	**Tỷ lệ%**
Anh - Việt	23,1	24,3	$+1{,}2_{(p<0{,}05)}$	5,31%

3.4 Kết luận chương

Trong chương này, chúng tôi đã trình bày về gióng hàng từ cho SMT. Chúng tôi đã đề xuất một số cải tiến mô hình IBM 1 theo cách tiếp cận dựa trên ràng buộc, cụ thể là: ràng buộc neo, ràng buộc về vị trí của từ, ràng buộc về từ loại và ràng buộc về cụm từ. Các ràng buộc này được sử dụng để hạn chế các gióng hàng không mong muốn giữa các từ trong một câu song ngữ, điều này không có được trong các mô hình IBM gốc. Với mỗi ràng buộc, chúng tôi đưa ra phương pháp tổng quát để tích hợp nó vào thuật toán EM trong quá trình ước lượng tham số của mô hình. Ngoài ra, chúng tôi đưa ra một phương pháp để kết hợp các ràng buộc. Các kết quả thực nghiệm cho thấy những cải tiến của chúng tôi đã nâng cao chất lượng dịch cho hệ thống SMT Anh - Việt.

Ràng buộc neo tạo ra một gióng hàng tin cậy giữa hai từ trong một câu song ngữ. Gióng hàng giữa hai từ trong một điểm neo được tạo ra bằng cách thiết lập xác suất gióng hàng bằng không ở vị trí đó cho tất cả các từ khác. Chúng tôi đã sử dụng các *cognate* và các cặp từ vựng có xác suất cao từ tập dữ liệu huấn luyện để làm điểm neo. Đây là cách làm đơn giản nhưng khá hiệu quả trong việc cải thiện chất lượng dịch cho SMT. Trong khi đó, các ràng buộc về vị trí của từ, ràng buộc về từ loại và ràng buộc về cụm từ giúp thu hẹp phạm vi gióng hàng giữa hai từ. Ràng buộc về vị trí của từ giới hạn phạm vi gióng hàng giữa các từ trong một câu song ngữ. Với mỗi cặp từ (trong câu song ngữ), chúng tôi gán trọng số cao hơn nếu ràng buộc về vị trí của từ được thỏa mãn và trọng số thấp hơn trong

trường hợp ngược lại. Ràng buộc về từ loại đòi hỏi mỗi từ trong câu nguồn chỉ gióng hàng với các từ trong câu đích có cùng quan hệ về POS. Ràng buộc về cụm từ yêu cầu mỗi từ trong cụm từ nguồn chỉ gióng hàng với các từ trong cụm từ đích. Các cụm từ được xác định bằng cách sử dụng các mẫu cú pháp.

Từ các thực nghiệm và quan sát, chúng tôi thấy có một số ngoại lệ. Cụ thể, một số cặp từ không thỏa mãn ràng buộc nhưng nó là dịch của nhau hoặc ngược lại. Điều này thực tế không ảnh hưởng nhiều đến kết quả toàn cục, bởi vì ở đây chúng tôi chỉ xét các ràng buộc với mỗi cặp câu trên một ngữ liệu song ngữ lớn. Vì thế, số trường hợp xảy ra ngoại lệ rất nhỏ so với các trường hợp đúng. Tất nhiên, việc giảm các trường hợp ngoại lệ sẽ làm tăng thêm hiệu quả của việc sử dụng các ràng buộc này. Trong tương lai, chúng tôi sẽ nghiên cứu các phương pháp xử lý riêng cho các trường hợp ngoại lệ.

Phương pháp chúng tôi trình bày tổng quát vì thế chúng ta có thể mở rộng để thêm ràng buộc mới. Chúng ta có thể sử dụng riêng lẻ hoặc kết hợp các ràng buộc lại với nhau như chúng tôi đã làm. Chúng tôi nghĩ rằng trong một số trường hợp, các ràng buộc có thể bị loại trừ lẫn nhau. Tức là, khi ràng buộc này thỏa mãn có thể ràng buộc kia lại không thỏa mãn. Điều này có thể ảnh hưởng đến chất lượng của gióng hàng khi ta áp dụng nhiều ràng buộc đồng thời. Do đó, việc khảo sát và lựa chọn ràng buộc tối ưu để sử dụng chúng vào việc cải tiến gióng hàng từ cho SMT là một bài toán có ý nghĩa đáng để nghiên cứu.

Chương 4

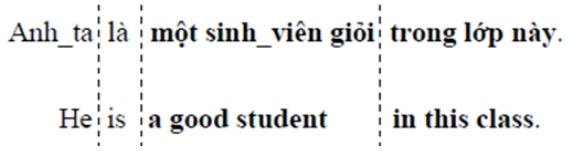
Xác định cụm từ song ngữ cho dịch máy thống kê

Trong chương này, chúng tôi trình bày việc xác định cụm từ song ngữ cho SMT. Chúng tôi đề xuất phương pháp sử dụng các mẫu cú pháp kết hợp với gióng hàng cụm từ để xác định cụm từ song ngữ. Các cụm từ song ngữ này đã được ứng dụng vào việc nâng cao chất lượng dịch cho hệ thống SMT Anh - Việt. Các thực nghiệm và đánh giá được trình bày ở cuối chương.

4.1 Bài toán rút trích cụm từ song ngữ

Trong ngôn ngữ học, cụm từ là một nhóm từ (hoặc đôi khi là một từ duy nhất) tạo thành một thành phần và có chức năng như một đơn vị duy nhất trong cú pháp của câu. Giả sử chúng ta đang làm việc với hai ngôn ngữ, tiếng Anh và tiếng Việt. Cho một cụm từ pe ở ngôn ngữ nguồn (tiếng Anh) và một cụm từ pv ở ngôn ngữ đích (tiếng Việt). Chúng tôi định nghĩa một cặp cụm từ $p = (pe, pv)$ là một cụm từ song ngữ nếu cụm từ nguồn pe và cụm từ đích pv là bản dịch của nhau, tức là, không có bổ sung từ trong cụm từ đích mà không thể tìm thấy từ tương ứng trong cụm từ nguồn và ngược lại [99].

Hình 4.1 cho thấy một ví dụ về các cụm từ song ngữ trong một câu song ngữ Anh - Việt. Chúng ta có thể thấy trong hình, có hai cụm từ: *"a good student"* và *"in this class"* trong câu tiếng Anh. Tương tự như vậy, câu tiếng Việt chứa hai



Hình 4.1: Ví dụ về các cụm từ song ngữ trong một câu song ngữ Anh - Việt, các từ in đậm chỉ ra các cụm từ.

cụm từ: *"một sinh_ viên giỏi"* và *"trong lớp này"*. Ở đây, các cụm từ song ngữ sẽ là: (*"một sinh_ viên giỏi"*,*"a good student"*) và (*"trong lớp này"*,*"in this class"*).

Cho ngữ liệu $C = \{(\mathbf{f}^{(l)}, \mathbf{e}^{(l)})\}$ chứa các câu song ngữ Anh - Việt. Trong đó, $1 \leq l \leq N$ và N là kích thước của ngữ liệu. Bài toán đặt ra ở đây là tìm và rút trích các cụm từ song ngữ trong ngữ liệu C. Lưu ý rằng, ở đây chúng tôi chỉ rút trích các cụm từ song ngữ chứa nhiều hơn một từ.

4.2 Phương pháp rút trích cụm từ song ngữ

Trong phần này, chúng tôi trình bày các bước để rút trích cụm từ song ngữ, sử dụng các mẫu cú pháp kết hợp với gióng hàng cụm từ, bao gồm: xác định cụm từ, tìm cụm từ đích và rút trích cụm từ song ngữ.

4.2.1 Xác định cụm

Chúng tôi sử dụng các mẫu cú pháp song ngữ được xác định trước để phát hiện và rút trích các cụm từ song ngữ từ ngữ liệu song ngữ Anh - Việt. Trong nghiên cứu này, chúng tôi thiết kế các mẫu cú pháp như là các chuỗi POS. Các mẫu cú pháp này được sử dụng để xác định cụm từ. Bảng 4.1 mô tả một số ví dụ về mẫu cú pháp và cụm từ tương ứng trong tiếng Anh.

Giả sử chúng ta có một cặp câu (**f**,**e**) từ ngữ liệu song ngữ so khớp với một cặp mẫu cú pháp tại các vị trí (j_1, j_2) trong câu nguồn và (i_1, i_2) trong câu đích. Từ đó, chúng ta dễ dàng rút trích các cụm từ nguồn $pe = f_{j_1}...f_{j_2}$ và cụm từ đích $pv = e_{i_1}...e_{i_2}$ (ở đây (pe, pv) là cặp ứng viên của cụm từ song ngữ). Tuy nhiên, do sự khác biệt về cấu trúc ngữ pháp giữa ngôn ngữ nguồn và ngôn ngữ đích cùng với quá trình gán nhãn từ loại cho văn bản tại mỗi ngôn ngữ có thể xảy ra lỗi. Những điều này sẽ làm giảm số cụm từ song ngữ được tìm thấy khi ta thực hiện

việc so sánh các mẫu cú pháp ở cả hai phía (câu nguồn và câu đích). Vì vậy, trong trường hợp chỉ so khớp ở một phía (trong câu **f** hoặc **e**), chúng tôi xác định cụm từ này (chúng tôi gọi là cụm từ nguồn) và tìm cụm từ còn lại (chúng tôi gọi là cụm từ đích).

Chúng ta có thể thấy trong Hình 4.1, một so khớp của mẫu cú pháp "DT/JJ/NN" được tìm thấy. Như vậy, cụm từ nguồn $pe =$ *"a good student"* sẽ được phát hiện và rút trích.

BẢNG 4.1: Một số ví dụ về mẫu cú pháp và cụm từ tương ứng trong tiếng Anh.

STT	**Mẫu cú pháp**	**Cụm từ**
1.	DT/NN	a book
		this computer
2.	DT/NNS	the books
		these employees
3.	DT/JJ/NN	that interesting book
		a good student

4.2.2 Tìm cụm từ đích

Giả sử chúng ta đã xác định được cụm từ nguồn $pe = f_{j_1}...f_{j_2}$ ở trong câu **f**, bây giờ chúng ta cần tìm một chuỗi các từ $e_{i_1}...e_{i_2}$ trong câu **e**, là bản dịch của cụm từ nguồn. Để thực hiện công việc này, chúng tôi sử dụng phương pháp gióng hàng cụm từ của Vogel [117] được trình bày trong công thức (4.1).

$$\begin{aligned} Pr_{i_1,i_2}(\mathbf{f}|\mathbf{e}) = & \prod_{j=1}^{j_1-1} \sum_{i \notin (i_1..i_2)} \frac{1}{I-k} t(f_j|e_i) \\ & \times \prod_{j=j_1}^{j_2} \sum_{i=i_1}^{i2} \frac{1}{k} t(f_j|e_i) \\ & \times \prod_{j=j_2+1}^{J} \sum_{i \notin (i_1..i_2)} \frac{1}{I-k} t(f_j|e_i) \end{aligned} \tag{4.1}$$

Ranh giới i_1 và i_2 của cụm từ pv trong câu đích được xác định bởi công thức (4.2).

$$(i_1, i_2) = \arg\max_{i_1,i_2} \{Pr_{i_1,i_2}(\mathbf{f}|\mathbf{e})\} \tag{4.2}$$

Trong công thức (4.1), $t(f_j|e_i)$ là xác suất dịch từ vựng (xác suất từ e_i dịch sang từ f_j). Chúng tôi sử dụng Thuật toán 3.1 (ở Chương 3) để tính xác suất này.

4.2.3 Rút trích cụm từ

Bây giờ, chúng tôi thực hiện rút trích các ứng viên của cụm từ song ngữ, như sau:

- Tính xác suất $t(f|e)$
- Với mỗi cặp câu $(\mathbf{f}^{(l)}, \mathbf{e}^{(l)}), 1 \leq l \leq N$[1]:
  - Với mỗi cặp mẫu cú pháp trong tập các mẫu cú pháp được xác định trước:
    - Nếu một cặp mẫu cú pháp được so khớp thì (pe, pv) là một ứng viên của cụm từ song ngữ.
    - Ngoài ra, nếu một mẫu cú pháp trong ngôn ngữ nguồn được so khớp thì rút trích cụm từ nguồn pe và tìm kiếm cụm từ đích pv dùng công thức (4.2).

Tiếp theo, để lọc cụm từ song ngữ (loại bỏ các cụm sai), chúng tôi tính xác suất dịch cụm từ bằng cách sử dụng tần suất tương đối:

$$Pr(pv|pe) = \frac{N(pv, pe)}{N(pe)} \tag{4.3}$$

Trong công thức (4.3), pe và pv lần lượt là cụm từ nguồn và đích. $N(pe, pv)$ là số lần cụm pe được dịch bởi pv và $N(pe)$ là số lần pe xuất hiện trong ngữ liệu. Lambert và cộng sự [94] đã chỉ ra rằng dữ liệu thưa có thể làm cho xác suất ước lượng theo cách này được đánh giá quá cao. Vì thế, và xác suất nghịch đảo $Pr(pe|pv)$ đã được chứng minh sẽ ước lượng tốt hơn. Để tăng độ tin cậy, chúng tôi sử dụng giá trị nhỏ nhất của hai tần suất tương đối như là xác suất dịch cụm từ, như thể hiện trong công thức (4.4).

$$Pr'(pv|pe) = \min(Pr(pv|pe), Pr(pe|pv)) \tag{4.4}$$

[1] N là kích thước của ngữ liệu.

Thuật toán 4.1 mô tả phương pháp của chúng tôi đề xuất để giải quyết bài toán rút trích cụm từ song ngữ. Đầu vào cho thuật toán là một tập hợp các cặp câu (**f**,**e**) và tập S chứa các cặp mẫu cú pháp $s = (se, sv)$ được xác định trước.

Thuật toán 4.1 Rút trích cụm từ song ngữ.

Đầu vào: Tập các câu song ngữ (**f**,**e**), S là tập các mẫu cú pháp $s = (se, sv)$.
Đầu ra: B là tập hợp các cụm từ song ngữ

```
1: tính xác suất t(f|e)
2: //khởi tạo
3: B = ∅
4: for all các cặp câu (f,e) do
5:    //so khớp các mẫu cú pháp
6:    if cặp mẫu cú pháp s = (se, sv) ∈ S phù hợp (được so khớp) then
7:       rút trích ứng viên của cụm từ song ngữ p = (pe, pv)
8:       B = B ∪ {p}
9:    else if mẫu cú pháp se ∈ S (ở câu nguồn) phù hợp then
10:      rút trích cụm từ ở câu nguồn pe
11:      tìm cụm từ ở câu đích pv dùng công thức (4.2)
12:      B = B ∪ {p}
13:   end if
14: end for
15: //lọc các cụm từ song ngữ
16: for all các ứng viên của cụm từ song ngữ p = (pe, pv) do
17:   if Pr(pv|pe) < θ then
18:      B = B \ {p}
19:   end if
20: end for
```

4.3 Tích hợp cụm từ song ngữ vào dịch máy

Bảng cụm từ (hay còn gọi là *phrase table*) đóng vai trò rất quan trọng đối với các hệ thống SMT dựa trên cụm từ. Tuy nhiên, như Ren và cộng sự trong [99] đã chỉ ra rằng, do những lỗi trong quá trình gióng hàng từ tự động và sự mở rộng các từ không được gióng hàng (unaligned word) trong giai đoạn rút trích cụm từ, dẫn đến nhiều cụm từ vô nghĩa sẽ được rút trích và kết quả tính xác suất dịch cụm từ không đúng. Để giảm bớt vấn đề này, tương tự các nghiên cứu của Ren [99], Bouamor [9], chúng tôi tích hợp các cụm từ song ngữ sau khi được rút trích từ ngữ liệu vào hệ thống SMT Anh - Việt theo hai cách:

1. Xây dựng thêm một bảng cụm từ từ các cụm từ song ngữ được rút trích tự động. Xác suất dịch của các cụm được tính theo các công thức (4.3) và (4.4). Bảng 4.2 trình bày ví dụ về một số cụm từ song ngữ được rút trích tự động từ ngữ liệu song ngữ Anh - Việt sử dụng trong thực nghiệm. Chúng tôi kết hợp bảng cụm từ ban đầu (được tạo ra trong quá trình huấn luyện mô hình dịch) và bảng cụm từ được tạo ra từ các cụm từ song ngữ vào trong hệ thống SMT. Như vậy, đối với mỗi cụm từ trong câu đầu vào, trong quá trình dịch bộ giải mã sẽ tìm kiếm tất cả các cụm dịch ứng cử viên trong cả hai bảng cụm từ (chúng tôi gọi là phương pháp APT).

2. Sử dụng các cụm từ song ngữ được rút trích tự động như là câu song ngữ và thêm chúng vào dữ liệu huấn luyện, sau đó huấn luyện lại mô hình dịch. Bằng cách tăng số lần xuất hiện của các cụm từ song ngữ, đây là những cụm từ song ngữ có chất lượng tốt. Với cách làm này, các gióng hàng lỗi sẽ giảm và việc ước lượng xác suất dịch cụm từ sẽ hợp lý hơn [99] (chúng tôi gọi là phương pháp ABP).

Bảng 4.2: Ví dụ về một số cụm từ song ngữ được sử dụng trong thực nghiệm.

Cụm từ tiếng Anh	**Cụm từ tiếng Việt**	**Xác suất**
a useful contact	một đầu_mối hữu_ích	0.4999
gusts of rage	cơn giận điên lên	0.9999
an american military official	một quan_chức quân_đội mỹ	0.9999
a second language	một ngôn_ngữ thứ_hai	0.7999
a normal reaction	một phản_ứng thông_thường	0.9999
the right number	con_số chính_xác	0.3333
a pleasant surprise	một sự ngạc_nhiên thú_vị	0.3333
a good journey	chúc thượng_lộ_bình_an	0.3333
a cheap hotel	một khách_sạn rẻ_tiền	0.9999
a famous conductor	người chỉ_huy_dàn_nhạc nổi_tiếng	0.9999

4.4 Thực nghiệm

4.4.1 Thực nghiệm về rút trích cụm từ song ngữ

4.4.1.1 Cài đặt thực nghiệm

Để đánh giá hiệu quả của việc rút trích cụm từ song ngữ từ ngữ liệu song ngữ, chúng tôi sử dụng các độ đo *precision* và *recall* như sau:

$$Precision = \frac{|X \cap Y|}{|X|} \tag{4.5}$$

$$Recall = \frac{|X \cap Y|}{|Y|} \tag{4.6}$$

Trong đó,

- X là tập hợp các cụm từ song ngữ được rút trích theo phương pháp áp dụng.
- Y là tập hợp các cụm từ song ngữ có trong ngữ liệu.

Ngoài ra, để cân bằng giữa độ chính xác và độ bao phủ, chúng sử dụng độ đo F_{score} như sau:

$$F_{score} = \frac{(2 * Precision * Recall)}{(Precision + Recall)} \tag{4.7}$$

Các thực nghiệm về rút trích cụm từ song ngữ được thực hiện trên 5.000 câu song ngữ Anh - Việt (được tạo bởi Hoàng Cường và cộng sự [44]). Để gán nhãn từ loại cho dữ liệu thực nghiệm, chúng tôi sử dụng các bộ công cụ: *vnTagger* cho văn bản tiếng Việt và *posTagger-1.0* cho văn bản tiếng Anh. Chúng tôi xây dựng một tập hợp các cặp mẫu cú pháp tiếng Anh và tiếng Việt, tập này bao gồm 10 cặp mẫu. Bảng 4.3 mô tả thông tin về các cặp mẫu cú pháp này.

4.4.1.2 Kết quả thực nghiệm

Chúng tôi đã thử nghiệm với một số giá trị của ngưỡng θ. Chi tiết được tóm tắt trong Bảng 4.4. Chúng ta có thể thấy, điểm *precision* (độ chính xác) cao hơn nếu

Bảng 4.3: 10 mẫu cú pháp song ngữ Anh - Việt được sử dụng để xác định cụm từ cho SMT.

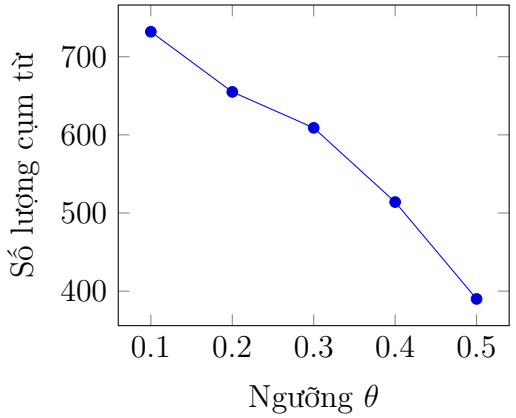
STT	**Tiếng Anh**	**Tiếng Việt**
1.	DT/NN	M/Nc/N, Nc/N/P
2.	DT/NNS	L/Nc/N/P, L/N/P
3.	DT/JJ/NN	Nc/N/A/P, M/Nc/N/A, M/N/A
4.	DT/JJ/JJ/NN	Nc/N/A/A/P, M/Nc/N/A/A, M/N/A/A
5.	DT/JJ/NNS	L/N/A
6.	DT/JJ/JJ/NNS	L/N/A/A
7.	PRP$/NN	Nc/N/E/P
8.	PRP$/JJ/NN	N/A/E/P
9.	PRP$/NNS	L/N/E/P
10.	PRP$/JJ/NNS	L/N/A/E/P

Bảng 4.4: Kết quả thử nghiệm sử dụng một số giá trị của ngưỡng θ.

Ngưỡng θ	**Precision**	**Recall**	F_score	**Số lượng cụm từ**
không sử dụng	75,97	13,25	22,57	828
0,05	79,36	13,04	22,40	780
0,10	82,24	12,68	21,98	732
0,15	83,71	12,24	21,36	694
0,20	84,73	11,69	20,55	655
0,25	**86,24**	**11,36**	**20,07**	**625**
0,50	87,94	7,22	13,36	390

giá trị của ngưỡng θ tăng lên và dĩ nhiên, giá trị của *recall* (độ bao phủ) sẽ giảm xuống. Cần lưu ý rằng trong bài toán này, độ chính xác là tiêu chí quan trọng nhất để đánh giá hiệu quả của hệ thống. Theo kết quả từ các thực nghiệm, chúng tôi thấy rằng với ngưỡng $\theta = 0,25$ chúng tôi đạt được kết quả tốt nhất, trong đó sự cân bằng giữa *precision* và *recall* được đảm bảo. Hình 4 cho thấy mối tương quan giữa ngưỡng θ và số lượng các cụm từ rút trích được. Số lượng các cụm từ giảm

xuống khi chúng ta tăng giá trị của ngưỡng θ. Chúng ta có thể thấy trong biểu đồ, khi $\theta = 0,1$ có 732 cụm từ song ngữ và con số này giảm xuống còn 655 khi $\theta = 0,2$. Bảng 4.5 trình bày kết quả thực nghiệm với phương pháp của chúng tôi



Hình 4.2: Tương quan giữa ngưỡng θ và số lượng cụm từ song ngữ.

và phương pháp so khớp mẫu cú pháp ở hai phía như trong [7]. Kết quả cho thấy, phương pháp của chúng tôi đạt được điểm số cao hơn trên cả hai độ đo *precision* và *recall*. Cụ thể, điểm F_{score} của phương pháp chúng tôi là $36,07$ trong khi của phương pháp so khớp mẫu cú pháp ở hai phía là $20,07$. Các kết quả này đã cho thấy phương pháp chúng tôi đề xuất là khá hiệu quả.

Bảng 4.5: Kết quả thực nghiệm với phương pháp của chúng tôi và phương pháp so khớp mẫu cú pháp ở hai phía.

Phương pháp	**Precision**	**Recall**	F_score	**Số lượng cụm từ**
So khớp mẫu cú pháp ở hai phía	86,24	11,36	20,07	625
Phương pháp chúng tôi đề xuất	89,12	22,61	36,07	1.204

4.4.2 Thực nghiệm về tích hợp cụm từ song ngữ vào dịch máy

4.4.2.1 Cài đặt thực nghiệm

Chúng tôi sử dụng 200.000 câu song ngữ Anh - Việt được tạo bởi Hoàng Cường và cộng sự [44]. Bảng 4.6 thống kê chi tiết về các thông số của ngữ liệu này. Hệ

thống SMT Anh - Việt dựa trên cụm từ được xây dựng với các thành phần như sau:

- Mô hình ngôn ngữ với công cụ SRILM: Chúng tôi xây dựng mô hình ngôn ngữ *3-gram* sử dụng kỹ thuật làm trơn Kneyser-Ney trên ngữ liệu 1.430.177 câu tiếng Việt chứa 22.056.253 từ và 317.028 từ vựng.
- Mô hình dịch và giải mã sử dụng công cụ MOSES [61].

Tập dữ liệu bao gồm 1.000 câu song ngữ Anh - Việt được sử dụng để đánh giá chất lượng dịch.

Bảng 4.6: Thống kê các thông số của ngữ liệu 200.000 câu song ngữ Anh - Việt được sử dụng trong thực nghiệm.

Ngôn ngữ	**Số từ**	**Số từ vựng**
Tiếng Anh	1.979.826	64.051
Tiếng Việt	2.028.514	51.246

Bảng 4.7: Thống kê về số lượng cụm từ song ngữ Anh - Việt được sử dụng trong thực nghiệm.

Kích thước ngữ liệu	**Số cụm từ**
100.000	16.327
200.000	33.791

4.4.2.2 Kết quả thực nghiệm

Bảng 4.8 tổng hợp các kết quả thực nghiệm khi tích hợp các cụm từ song ngữ được rút trích tự động từ hai tập ngữ liệu: 100.000 và 200.000 câu song ngữ vào hệ thống SMT Anh - Việt: (1) không xử lý cụm từ (baseline), (2) thêm cụm từ vào dữ liệu và huấn luyện lại mô hình, (3) xây dựng thêm một bảng cụm từ, (4) kết hợp giữa (2) và (3). Thống kê về số lượng cụm từ song ngữ sử dụng trong thực nghiệm được trình bày ở bảng 4.7. Trong bảng 4.8, các ký hiệu Δ_1, Δ_2, Δ_3 lần lượt là độ chênh lệnh điểm BLEU (tăng (+)/giảm (−)) của các phương pháp (2), (3) và (4) so với (1). Chúng ta có thể thấy, xử lý cụm từ trong SMT đạt được

chất lượng dịch tốt hơn so với không xử lý cụm từ. Cụ thể, chất lượng dịch tăng tương ứng là 0,35 và 0,41 điểm BLEU khi thêm cụm từ vào dữ liệu huấn luyện (sau đó huấn luyện lại mô hình dịch) và xây dựng thêm một bảng cụm từ (từ các cụm từ song ngữ được rút trích tự động). Ngoài ra, kết hợp giữa hai phương pháp đạt được kết quả cao hơn với điểm BLEU tăng 0,53.

Bảng 4.8: Kết quả thử nghiệm khi tích hợp các cụm từ song ngữ vào hệ thống SMT Anh - Việt.

Ngữ liệu	Baseline (1)	ABP (2)	Δ_1 (+/−)	APT (3)	Δ_2 (+/−)	ABP + APT (4)	Δ_3 (+/−)
100.000	25,73	26,05	+0,32	26,22	+0,49	26,27	+0,54
200.000	28,53	28,91	+0,38	28,87	+0,34	29,06	+0,53

4.5 Kết luận chương

Chúng tôi đã trình bày phương pháp dựa trên cách tiếp cận lai để rút trích cụm từ song ngữ từ ngữ liệu song ngữ Anh - Việt và ứng dụng cho SMT. Phương pháp của chúng tôi kết hợp giữa các mẫu cú pháp được xác định trước và xác suất dịch cụm từ để rút trích các cụm từ song ngữ. Bằng cách sử dụng các mẫu cú pháp và kết hợp với gióng hàng cụm từ để tìm bản dịch của cụm từ nguồn. Vì thế, chúng tôi đã rút trích nhiều cụm từ song ngữ hơn. Các kết quả đạt được đã cho thấy hiệu quả của đề xuất này. Khi tích hợp các cụm từ song ngữ được rút trích tự động vào hệ thống SMT, chất lượng dịch đã cải thiện đáng kể. Trong tương lai, chúng tôi sẽ mở rộng phương pháp này theo một số hướng. Đầu tiên, chúng tôi sẽ sử dụng nhiều hơn nữa các mẫu cú pháp để tăng số lượng các cụm từ được rút trích. Thứ hai, chúng tôi dự định sử dụng phương pháp được đề xuất để xây dựng một từ điển cụm từ song ngữ cho cặp ngôn ngữ Anh - Việt và sẽ cải thiện chất lượng của SMT Anh - Việt bằng cách sử dụng từ điển này.

Kết luận

Trong phần này, chúng tôi tóm lược lại các kết quả chính và những đóng góp của luận án. Ngoài ra, chúng tôi trình bày một số hạn chế của luận án và thảo luận về hướng phát triển cho các nghiên cứu tiếp theo trong tương lai.

1. Tóm lược các kết quả và đóng góp của luận án

Luận án chúng tôi tập trung vào việc khai phá tri thức song ngữ và ứng dụng trong MT Anh - Việt. Chúng tôi đã đề xuất một số phương pháp để xây dựng ngữ liệu song ngữ cho SMT, đưa ra một số cải tiến mô hình IBM 1 để gióng hàng từ cho SMT và xác định cụm từ song ngữ cho SMT. Trong 4 chương của luận án, ngoài Chương 1 trình bày tổng quan về các vấn đề nghiên cứu trong luận án; nội dung và kết quả nghiên cứu được trình bày ở các chương chính là 2, 3 và 4. Các đóng góp chính và kết quả của luận án có thể được tóm tắt như sau:

Thứ nhất, chúng tôi đã đề xuất một số phương pháp để xây dựng ngữ liệu song ngữ cho SMT. Cụ thể, chúng tôi khai thác từ hai nguồn: Web và sách điện tử song ngữ. Đối với nguồn từ Web, chúng tôi rút trích các văn bản song ngữ từ các trang web song ngữ Anh - Việt; đưa ra hai phương pháp thiết kế các đặc trưng dựa trên nội dung: dựa trên *cognate* và dựa trên việc xác định các phân đoạn dịch. Sau đó, chúng tôi kết hợp các đặc trưng dựa trên nội dung với các đặc trưng dựa trên cấu trúc và mô hình hóa bài toán này như bài toán phân loại để trích rút các văn bản song ngữ. Các phương pháp chúng tôi đề xuất đạt được kết quả tốt hơn (độ chính xác 88,2% và 90,0%) so với phương pháp sử dụng các đặc trưng dựa vào cấu trúc trang *web* (độ chính xác 44,4%) và phương pháp sử dụng từ điển (độ chính xác 65,2%). Đối với nguồn từ sách điện tử song ngữ, chúng tôi sử dụng một số mẫu liên kết giữa các khối văn bản trong hai ngôn ngữ để rút trích các câu song ngữ. Các thực nghiệm về rút trích câu song ngữ từ sách điện tử theo phương pháp chúng tôi đề xuất đã đạt được $95,0\%$ theo độ đo F_{score}.

Thứ hai, chúng tôi đã đề xuất một số cải tiến đối với mô hình IBM 1 theo cách tiếp cận dựa trên ràng buộc, bao gồm: ràng buộc neo, ràng buộc về vị trí của từ, ràng buộc về từ loại và ràng buộc về cụm từ. Với mỗi ràng buộc, chúng tôi đưa ra phương pháp tổng quát để tích hợp nó vào thuật toán EM trong quá trình ước lượng tham số của mô hình. Ngoài ra, chúng tôi đưa ra một phương pháp để kết hợp các ràng buộc. Việc cải tiến này giúp nâng cao chất lượng dịch cho các

hệ thống SMT. Cụ thể, với phương pháp kết hợp ràng buộc, chất lượng MT tăng 7,16% so với mô hình IBM gốc và tăng 5,31% so với sử dụng Giza++.

Thứ ba, chúng tôi đã đề xuất phương pháp rút trích cụm từ song ngữ từ ngữ liệu song ngữ, sử dụng các mẫu cú pháp kết hợp với gióng hàng cụm từ. Các cụm từ song ngữ này được ứng dụng vào việc tăng chất lượng SMT. Các thực nghiệm được thực hiện trên hệ thống SMT Anh - Việt cho thấy phương pháp xác định cụm từ song ngữ như chúng tôi đưa ra đạt được chất lượng dịch tốt hơn so với không xử lý cụm từ, cụ thể trong trường hợp tốt nhất điểm BLEU tăng 0,53.

2. Hạn chế và hướng phát triển của luận án

Chúng tôi đã nghiên cứu, đề xuất một số phương pháp để khai thác tri thức song ngữ và ứng dụng trong MT Anh - Việt. Chúng tôi hy vọng rằng, đây là bước khởi đầu cho các nghiên cứu tiếp theo về SMT nói chung và SMT cho cặp ngôn ngữ Anh - Việt nói riêng. Trong quá trình nghiên cứu và thực nghiệm, chúng tôi nhận thấy một số hạn chế và hướng phát triển của luận án, cụ thể như sau:

Thứ nhất, đối với việc xây dựng ngữ liệu song ngữ, chúng tôi dự định sẽ tiếp tục khai thác nguồn từ Web với các thành phần song ngữ khác như đoạn, câu và cụm từ. Công việc này cũng sẽ rất có ý nghĩa trong trường hợp chất lượng dịch giữa các trang *web* song ngữ không tốt. Ngoài ra, với việc kết hợp nguồn từ Web và sách điện tử song ngữ theo các phương pháp đã đề xuất, chúng tôi sẽ thu thập ngữ liệu song ngữ cho cặp ngôn ngữ Anh - Việt đủ cho nhiều nhiệm vụ của NLP.

Thứ hai, với bài toán gióng hàng từ cho SMT, chúng tôi sẽ xử lý các ngoại lệ xảy ra đối với các ràng buộc như đã đề xuất. Việc giảm các trường hợp ngoại lệ sẽ làm tăng thêm hiệu quả của việc sử dụng các ràng buộc này. Vì thế, trong tương lai, chúng tôi sẽ nghiên cứu các phương pháp xử lý riêng cho các trường hợp ngoại lệ. Một hướng phát triển khác là chúng ta có thể thêm ràng buộc mới với cách làm tương tự, bởi vì các phương pháp chúng tôi đưa ra là tổng quát. Xa hơn nữa, chúng ta có thể kết hợp các ràng buộc theo nhiều cách khác nhau. Trong luận án này, chúng tôi đã đưa ra một cách để kết hợp chúng, dùng "phép hợp" giữa các ràng buộc. Chúng tôi nghĩ rằng trong một số trường hợp, các ràng buộc có thể bị loại trừ lẫn nhau. Tức là, khi ràng buộc này thỏa mãn có thể ràng buộc kia lại không thỏa mãn. Điều này có thể ảnh hưởng đến chất lượng của gióng hàng khi ta áp dụng nhiều ràng buộc đồng thời. Do đó, việc khảo sát và lựa chọn ràng buộc

tối ưu để sử dụng chúng vào việc cải tiến gióng hàng từ cho SMT cũng là một bài toán thú vị đáng để nghiên cứu.

Thứ ba, đối với bài toán xác định cụm từ song ngữ cho SMT, chúng tôi sẽ mở rộng phương pháp này theo một số hướng. Đầu tiên, chúng tôi sẽ sử dụng nhiều hơn nữa các mẫu cú pháp để tăng số lượng các cụm từ được rút trích. Thứ hai, chúng tôi dự định sử dụng phương pháp đã đề xuất để xây dựng một từ điển cụm từ song ngữ cho cặp ngôn ngữ Anh-Việt và sẽ cải thiện chất lượng của SMT Anh-Việt bằng cách sử dụng từ điển này.

Danh mục công trình khoa học của tác giả liên quan đến luận án

[1] Le Quang Hung and Le Anh Cuong (2010), "Extracting parallel texts from the web", *Proceedings of the Second International Conference on Knowledge and Systems Engineering, IEEE Computer Society*, pages 147-151.

[2] Le Quang Hung and Le Anh Cuong (2012), "Improving Word Alignment for Statistical Machine Translation Based on Constraints", *Asian Language Processing (IALP), International Conference on, IEEE Computer Society*, pages 113-116.

[3] Le Quang Hung and Le Anh Cuong (2012), "Statistical Word Alignment with Part-of-Speech Constraint", *Kỷ yếu hội thảo Quốc gia lần thứ XV "Một số vấn đề chọn lọc của Công nghệ thông tin và Truyền thông"*, trang 410-416.

[4] Quang-Hung LE, Duy-Cuong NGUYEN, Duc-Hong PHAM, Anh-Cuong LE, and Van-Nam HUYNH (2013), "Paragraph Alignment for English-Vietnamese Parallel E-Books", In *Knowledge and Systems Engineering, Springer International Publishing*, pages 251-259.

[5] Quang-Hung LE, Anh-Cuong LE, and Van-Nam HUYNH (2013), "Parallel phrase extraction from English-Vietnamese parallel corpora", In *Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RIVF), 2013 IEEE RIVF International Conference on*, pages 175-179.

[6] Le Quang Hung and Le Anh Cuong (2013), "An effective method to sentence alignment for the English-Vietnamese parallel e-book", *Kỷ yếu hội thảo Quốc gia lần thứ XVI "Một số vấn đề chọn lọc của Công nghệ thông tin và Truyền thông"*, trang 12-16.

[7] Le Quang Hung (2014), "A new approach to extract parallel corpus", *Tạp chí khoa học Trường Đại học Quy Nhơn*, Số 4, Tập VIII, trang 12-24.

[8] Quang-Hung LE and Anh-Cuong LE (2014), "Syntactic pattern based Word Alignment for Statistical Machine Translation", *The International Journal of Knowledge and Systems Science (IJKSS), IGI Global Publishing*, Volume 5 Issue 3, pages 36-45.

Tài liệu tham khảo

[1] Acosta, O., Villavicencio, A., and Moreira, V. (2011). Identification and treatment of multiword expressions applied to information retrieval. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 101–109, Portland, Oregon, USA. Association for Computational Linguistics.

[2] Attia, M., Toral, A., Tounsi, L., Pecina, P., and van Genabith, J. (2010). Automatic extraction of arabic multiword expressions. In *Proceedings of the Workshop on Multiword Expressions: from Theory to Applications (MWE 2010)*, pages 18–26, Beijing, China. Association for Computational Linguistics.

[3] Attia, M. A. (2006). Accommodating multiword expressions in an arabic lfg grammar. In *Proceedings of the 5th international conference on Advances in Natural Language Processing*, FinTAL'06, pages 87–98, Berlin, Heidelberg. Springer-Verlag.

[4] Ayan, N. F. (2005). *Combining linguistic and machine learning techniques for word alignment improvement*. PhD thesis, College Park, MD, USA.

[5] Bai, M.-H., You, J.-M., Chen, K.-J., and Chang, J. S. (2009). Acquiring translation equivalences of multiword expressions by normalized correlation frequencies. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2*, EMNLP '09, pages 478–486, Stroudsburg, PA, USA. Association for Computational Linguistics.

[6] Ban, D. Q. (2007). *Ngữ pháp tiếng Việt (tập 1)*. Nhà xuất bản Giáo dục.

[7] Baobao, C., Danielsson, P., and Teubert, W. (2002). Extraction of translation unit from chinese-english parallel corpora. In *Proceedings of the first SIGHAN workshop on Chinese language processing - Volume 18*, SIGHAN '02, pages 1–5, Stroudsburg, PA, USA. Association for Computational Linguistics.

[8] Berg-Kirkpatrick, T., Bouchard-Côté, A., DeNero, J., and Klein, D. (2010). Painless unsupervised learning with features. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 582–590. Association for Computational Linguistics.

[9] Bouamor, D., Semmar, N., and Zweigenbaum, P. (2012). Identifying bilingual multi-word expressions for statistical machine translation. In *LREC*, pages 674–679.

[10] Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Mercer, R., and Roosin, P. (1990). A statistical approach to machine translation. *Computational Linguistics*, pages 79–85.

[11] Brown, P. F., Lai, J. C., and Mercer, R. L. (1991). Aligning sentences in parallel corpora. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, ACL '91, pages 169–176, Stroudsburg, PA, USA. Association for Computational Linguistics.

[12] Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: parameter estimation. *Comput. Linguist.*, 19(2):263–311.

[13] Brunning, J. J. J. (2010). *Alignment Models and Algorithms for Statistical Machine Translation*. PhD thesis, University of Cambridge.

[14] Cambazoglu, B. B., Karaca, E., Kucukyilmaz, T., Turk, A., and Aykanat, C. (2007). Architecture of a grid-enabled web search engine. *Information Processing and Management*, pages 609–623.

[15] Charitakis, K. (2007). Using parallel corpora to create a greek-english dictionary with uplug. In *Proc. 16th Nordic Conference on Computational Linguistics-NODALIDA '07*.

[16] Chen, J., Chau, R., and Yeh, C.-H. (2004). Discovering parallel text from the world wide web. *In Proceedings Australasian Workshop on Data Mining and Web Intelligence (DMWI)*, pages 157–161.

[17] Chen, J. and J.Y., N. (2000). Automatic construction of parallel english-chinese corpus for cross-language information retrieval. *In Proceedings ANLP, Seattle*, pages 21–28.

[18] Chen, S. F. (1993). Aligning sentences in bilingual corpora using lexical information. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, ACL '93, pages 9–16, Stroudsburg, PA, USA. Association for Computational Linguistics.

[19] Clark, J. H., Dyer, C., Lavie, A., and Smith, N. A. (2011). Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 176–181. Association for Computational Linguistics.

[20] Clifton, A. and Sarkar, A. (2011). Combining morpheme-based machine translation with post-processing morpheme prediction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 32–42, Stroudsburg, PA, USA. Association for Computational Linguistics.

[21] Collier, N., Ono, K., and Hirakawa, H. (1998). An experiment in hybrid dictionary and statistical sentence alignment. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 268–274. Association for Computational Linguistics.

[22] Cowan, B., Kučerová, I., and Collins, M. (2006). A discriminative model for tree-to-tree translation. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 232–241. Association for Computational Linguistics.

[23] Cruys, T. v. d. and Villada Moirón, B. (2007). Lexico-semantic multiword expression extraction. *LOT Occasional Series*, 7:175–190.

[24] Dang, V. B. and Bao-Quoc, H. (2007). Automatic construction of english-vietnamese parallel corpus through web mining. *In Proceedings of 5th IEEE International Conference on Computer Science - Research, Innovation and Vision of the Future (RIVF), Hanoi, Vietnam.*

[25] Davis, M. W. and Dunning, T. E. (1995). A trec evaluation of query translation methods for multi-lingual text retrieval. In *Fourth Text Retrieval Conference*, pages 483–498.

[26] DellaPietra, S. and DellaPietra, V. (1994). Candide: a statistical machine translation system. In *Proceedings of the workshop on Human Language Technology*, pages 457–457. Association for Computational Linguistics.

[27] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, 39(1):1–38.

[28] Dhouha Bouamor, Nasredine Semmar, P. r. Z. (2012). Automatic construction of a multiword expressions bilingual lexicon: A statistical machine translation evaluation perspective. In *Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon*, COLING 2012, pages 95–108.

[29] Dien, D., Kiem, H., and Van Toan, N. (2001). Vietnamese word segmentation. In *NLPRS*, volume 1, pages 749–756.

[30] Dinh, D., Kiem, H., and Hovy, E. (2003). Btl: a hybrid model for english-vietnamese machine translation. In *Proceedings of the MT Summit IX*, pages 23–27.

[31] Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145. Morgan Kaufmann Publishers Inc.

[32] Dyer, C., Chahuneau, V., and Smith, N. A. (2013). A simple, fast, and effective reparameterization of ibm model 2. In *HLT-NAACL*, pages 644–648. Citeseer.

[33] Dyer, C., Clark, J., Lavie, A., and Smith, N. A. (2011). Unsupervised word alignment with arbitrary features. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 409–419. Association for Computational Linguistics.

[34] Frankenberg-Garcia, A. and Santos, D. (2003). Introducing compara: the portuguese-english parallel corpus. *Corpora in translator education*, pages 71–87.

[35] Gale, W. A. and Church, K. W. (1993). A program for aligning sentences in bilingual corpora. *Computational linguistics*, 19(1):75–102.

[36] Galley, M., Graehl, J., Knight, K., Marcu, D., DeNeefe, S., Wang, W., and Thayer, I. (2006). Scalable inference and training of context-rich syntactic translation models. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 961–968. Association for Computational Linguistics.

[37] Gelbukh, A., Sidorov, G., and Vera-Félix, J. A. (2006). Paragraph-level alignment of an english-spanish parallel corpus of fiction texts using bilingual dictionaries. In *Proceedings of the 9th international conference on Text, Speech and Dialogue*, TSD'06, pages 61–67, Berlin, Heidelberg. Springer-Verlag.

[38] Ghaffar, S. A. and Fakhr, M. W. (2011). English to arabic statistical machine translation system improvements using preprocessing and arabic morphology analysis. In *Proceedings of the 13th IASME/WSEAS international conference on Mathematical Methods and Computational Techniques in Electrical Engineering conference on Applied Computing*, ACC'11/MMACTEE'11, pages 94–98, Stevens Point, Wisconsin, USA. World Scientific and Engineering Academy and Society (WSEAS).

[39] Gimpel, K. (2012). *Discriminative Feature-Rich Modeling for Syntax-Based Machine Translation*. PhD thesis, Carnegie Mellon University.

[40] Gomis, M. E., Martínez, F. S., and Forcada, M. L. (2012). A simple approach to use bilingual information sources for word alignment. *Procesamiento del lenguaje natural*, 49:93–100.

[41] Gupta, A. and Pala, K. (2012). A generic and robust algorithm for paragraph alignment and its impact on sentence alignment in parallel corpora. pages 18–27.

[42] Helft, M. (2010). Google's computing power refines translation tool. *New York Times (March 8, 2010) A*, 1.

[43] Hùng, V. T. (2007). Phương pháp và công cụ đánh giá tự động các hệ thống dịch tự động trên mạng. *Tạp chí Khoa học và Công nghệ, Đại học Đà Nẵng*, 18(1):37–42.

[44] Hoang, C., Le, A.-C., Nguyen, P.-T., and Ho, T.-B. (2012a). Exploiting non-parallel corpora for statistical machine translation. In *RIVF*, pages 1–6. IEEE.

[45] Hoang, C., Le, C. A., and Pham, S. B. (2012b). A systematic comparison between various statistical alignment models for statistical english-vietnamese

phrase-based translation. In *Knowledge and Systems Engineering (KSE), 2012 Fourth International Conference on*, pages 143–150. IEEE.

[46] Huang, L., Knight, K., and Joshi, A. (2006). Statistical syntax-directed translation with extended domain of locality. In *Proceedings of AMTA*, volume 2006, pages 223–226.

[47] Huyên, N. T. M., Roussanaly, A., Vinh, H. T., et al. (2008). A hybrid approach to word segmentation of vietnamese texts. In *Language and Automata Theory and Applications*, pages 240–249. Springer.

[48] Đinh Điền (2003). Dịch tự động anh - việt dựa trên việc học luật chuyển đổi từ ngữ liệu song ngữ. In *Luận án tiến sĩ*. Trường Đại học Khoa học Tự nhiên – Đại học Quốc gia TP. Hồ Chí Minh.

[49] Đinh Điền and Quốc, H. B. (2008). Vấn đề về ranh giới từ trong ngữ liệu song ngữ anh - việt. pages 1–10.

[50] Ittycheriah, A. and Roukos, S. (2005). A maximum entropy word aligner for arabic-english machine translation. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 89–96, Stroudsburg, PA, USA. Association for Computational Linguistics.

[51] Jurafsky, D. and James, H. (2000). Speech and language processing an introduction to natural language processing, computational linguistics, and speech.

[52] Kamigaito, H., Watanabe, T., Takamura, H., and Okumura, M. (2014). Unsupervised word alignment using frequency constraint in posterior regularized EM. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 153–158.

[53] Kay, M. (1973). Automatic translation of natural languages. *Daedalus*, pages 217–230.

[54] Khalid Al Khatib, A. B. (2010). Automatic extraction of arabic multi-word terms. In *Proceedings of the International Multiconference on Computer Science and Information Technology*, pages 411–418.

[55] Khanh, P. N. (2009). An approach to automatically search for parallel texts scattering across websites.

[56] Kneser, R. and Ney, H. (1995). Improved backing-off for m-gram language modeling. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, volume 1, pages 181–184. IEEE.

[57] Knight, K. (1999). A statistical mt tutorial workbook. In *Prepared for the 1999 JHU Summer Workshop.*

[58] Koehn, P., H. H. (2007). Factored translation models. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning.*

[59] Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. *In MT Summit.*

[60] Koehn, P. (2009). *Statistical machine translation.* Cambridge University Press.

[61] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180. Association for Computational Linguistics.

[62] Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54. Association for Computational Linguistics.

[63] Kondrak, G., Marcu, D., and Knight, K. (2003a). Cognates can improve statistical translation models. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003–short papers-Volume 2*, pages 46–48. Association for Computational Linguistics.

[64] Kondrak, G., Marcu, D., and Knight, K. (2003b). Cognates can improve statistical translation models. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003–short papers - Volume 2*, NAACL-Short '03, pages 46–48, Stroudsburg, PA, USA. Association for Computational Linguistics.

[65] Kumano, A. and Hirakawa, H. (1994). Building an mt dictionary from parallel texts based on linguisitic and statistical information. *In Proceedings 15th COLING*, pages 76–81.

[66] Lavie, A., Probst, K., Peterson, E., Vogel, S., Levin, L., Llitjós, A. F., and Carbonell, J. G. (2004). A trainable transfer-based machine translation approach for languages with limited resources.

[67] Lee, J.-H., Lee, S.-W., Hong, G., Hwang, Y.-S., Kim, S.-B., and Rim, H.-C. (2010). A post-processing approach to statistical word alignment reflecting alignment tendency between part-of-speeches. In *Coling 2010: Posters*, pages 623–629, Beijing, China. Coling 2010 Organizing Committee.

[68] Li, P., Sun, M., and Xue, P. (2010). Fast-champollion: a fast and robust sentence alignment algorithm. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 710–718. Association for Computational Linguistics.

[69] Lin, D. and Cherry, C. (2003). Word alignment with cohesion constraint. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003–short papers - Volume 2*, NAACL-Short '03, pages 49–51, Stroudsburg, PA, USA. Association for Computational Linguistics.

[70] Liu, Y., Liu, Q., and Lin, S. (2005). Log-linear models for word alignment. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 459–466, Stroudsburg, PA, USA. Association for Computational Linguistics.

[71] Liu, Y., Liu, Q., and Lin, S. (2006). Tree-to-string alignment template for statistical machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 609–616. Association for Computational Linguistics.

[72] Liu, Y., Liu, Q., and Lin, S. (2010). Discriminative word alignment by linear modeling. *Comput. Linguist.*, 36(3):303–339.

[73] Liu, Y., Lü, Y., and Liu, Q. (2009). Improving tree-to-tree translation with packed forests. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 558–566. Association for Computational Linguistics.

[74] Liu, Y. and Sun, M. (2014). Contrastive unsupervised word alignment with non-local features. *arXiv preprint arXiv:1410.2082*.

[75] Loevinger, L., Burks, A. R., Burks, A. W., and Mollenhoff, C. R. (1989). The first electronic computer: The atanasoff story. *Jurimetrics J*, 29:359.

[76] Ma, X. and Mark, L. (1999). Bits: A method for bilingual text search over the web. *Machine Translation Summit VII*.

[77] Ma, Y., Ozdowska, S., Sun, Y., and Way, A. (2008). Improving word alignment using syntactic dependencies. In *Proceedings of the Second Workshop on Syntax and Structure in Statistical Translation*, SSST '08, pages 69–77.

[78] McEwan, C., Ounis, I., and Ruthven, I. (2002). Advances in information retrieval. *Springer*, pages 365–368.

[79] Mermer, C., Saraçlar, M., and Sarikaya, R. (2013). Improving statistical machine translation using bayesian word alignment and gibbs sampling. *IEEE Transactions on Audio, Speech and Language Processing*, 21(5):1090–1101.

[80] Meyers, A., Kosaka, M., and Grishman, R. (1998). A multilingual procedure for dictionary-based sentence alignment. In *Proceedings of the Third Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup*, AMTA '98, pages 187–198, London, UK, UK. Springer-Verlag.

[81] Mitamura, T., Nyberg, E. H., and Carbonell, J. G. (1991). An efficient interlingua translation system for multi-lingual document production.

[82] Moore, R. C. (2004). Improving ibm word-alignment model 1. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 518. Association for Computational Linguistics.

[83] Moore, R. C. (2005). A discriminative framework for bilingual word alignment. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 81–88, Stroudsburg, PA, USA. Association for Computational Linguistics.

[84] Munteanu, D. and Marcu, D. (2005). Improving machine translation performance by exploiting comparable corpora. *Computational Linguistics*, pages 477–504.

[85] Munteanu, D. and Marcu, D. (2006). Extracting parallel sub-sentential fragments from non-parallel corpora. *ACL*, pages 81–88.

[86] Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

[87] Nagao, M. (1984). A framework of a mechanical translation between japanese and english by analogy principle. *Artificial and human intelligence*, pages 351–354.

[88] Nhung, N. T. H. (2008). Sử dụng mô hình xác suất cho bài toán chuyển đổi trật tự từ trong dịch máy thống kê anh – việt dựa trên ngữ. In *Luận văn Thạc sĩ, chuyên ngành Khoa học máy tính*. Trường Đại học Khoa học Tự nhiên – Đại học Quốc gia TP. Hồ Chí Minh.

[89] N.Westerhout, E. (2005). A corpus of dutch aphasic speech: Sketching the design and performing a pilot study.

[90] Oard, D. W. (1997). Cross-language text retrieval research in the usa. *Third DELOS Workshop, European Research Consortium for Informatics and Mathematics*.

[91] Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.

[92] Och, F. J., Ney, H., Josef, F., and Ney, O. H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29.

[93] Papineni, Kishore, Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. *ACL, Philadelphia*, pages 311–318.

[94] Patrik Lambert, R. B. (2005). Data inferred multi-word expressions for statistical machine translation. *Proceedings of Machine Translation Summit X*, pages 396–403.

[95] Pecina, P., Toral, A., Papavassiliou, V., Prokopidis, P., Tamchyna, A., Way, A., and van Genabith, J. (2015). Domain adaptation of statistical machine translation with domain-focused web crawling. *Language Resources and Evaluation*, 49(1):147–193.

[96] Špela Vintar and Fišer, D. (2008). Harvesting multi-word expressions from parallel corpora. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

[97] P.Resnik and Philip (1999). Mining the web for bilingual text. *In Proceedings of the 37th Annual Meeting of the ACL, College Park, MD*, pages 527–534.

[98] Rasooli, M. S., Kashefi, O., and Minaei-Bidgoli, B. (2011). Extracting parallel paragraphs and sentences from english-persian translated documents. In *Information Retrieval Technology*, pages 574–583. Springer.

[99] Ren, Z., Lü, Y., Cao, J., Liu, Q., and Huang, Y. (2009). Improving statistical machine translation using domain bilingual multiword expressions. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, MWE '09, pages 47–54, Stroudsburg, PA, USA. Association for Computational Linguistics.

[100] Resnik, P. and Philip (1998). Parallel strands: A preliminary investigation into mining the web for bilingual text. *In Proceedings of the Third Conference of the Association for Machine Translation in the Americas (AMTA). Langhorne, PA*, pages 28–31.

[101] Resnik, P. and Smith, N. A. (2003). The web as a parallel corpus. *Computational Linguistics*, pages 349–380.

[102] SanJuan, E. and Ibekwe-SanJuan, F. (2006). Text mining without document context. *Inf. Process. Manage.*, 42(6):1532–1552.

[103] Sato, S. and Nagao, M. (1990). Toward memory-based translation. In *Proceedings of the 13th conference on Computational linguistics-Volume 3*, pages 247–252. Association for Computational Linguistics.

[104] Sellami, R., Deffaf, F., Sadat, F., and Hadrich Belguith, L. (2015). Improved statistical machine translation by cross-linguistic projection of named entities recognition and translation. *Computación y Sistemas*, 19(4).

[105] Sennrich, R. and Volk, M. (2010). Mt-based sentence alignment for ocr-generated parallel texts. In *The Ninth Conference of the Association for Machine Translation in the Americas (AMTA 2010), Denver, Colorado.*

[106] Sennrich, R. and Volk, M. (2011). Iterative, mt-based sentence alignment of parallel texts.

[107] Shen, L., Xu, J., and Weischedel, R. M. (2008). A new string-to-dependency machine translation algorithm with a target dependency language model. In *ACL*, pages 577–585. Citeseer.

[108] Siham Boulaknadel, B. D. and Aboutajdine, D. (2008). A multi-word term extraction program for arabic language. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

[109] Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Weischedel, R. (2006). A study of translation error rate with targeted human annotation. In *In Proceedings of the Association for Machine Transaltion in the Americas (AMTA 2006.*

[110] Songyot, T. and Chiang, D. (2014). Improving word alignment using word similarity. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1840–1845.

[111] Talbot, D. (2005). Constrained em for parallel text alignment. *Nat. Lang. Eng.*, 11(3):263–277.

[112] Tamura, A., Watanabe, T., and Sumita, E. (2014). Recurrent neural networks for word alignment model. In *Proc. ACL*, pages 1470–1480.

[113] Taskar, B., Lacoste-Julien, S., and Klein, D. (2005). A discriminative matching approach to word alignment. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 73–80, Stroudsburg, PA, USA. Association for Computational Linguistics.

[114] Tay, R. and Ibrahim, T. (2010). Research on paragraph alignment technology in chinese-uighur bilingual corpus. *Journal of Xinjiang University (Natural Science Edition)*, 1:021.

[115] Varea, I. G., Och, F. J., Ney, H., and Casacuberta, F. (2002). Improving alignment quality in statistical machine translation using context-dependent maximum entropy models. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics.

[116] Vaswani, A., Huang, L., and Chiang, D. (2012). Smaller alignment models for better translations: unsupervised word alignment with the l 0-norm. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 311–319. Association for Computational Linguistics.

[117] Vogel, S. (2005). Pesa: Phrase pair extraction as sentence splitting. In *in Proceedings: the tenth Machine Translation*.

[118] Volk, M., Vintar, S., and Buitelaar, P. (2003). Ontologies in cross-language information retrieval. *In Proceedings of WOW2003*, pages 43–50.

[119] Xu, J. and Chen, J. (2011). How much can we gain from supervised word alignment? In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 165–169. Association for Computational Linguistics.

[120] Yamada, K. and Knight, K. (2001). A syntax-based statistical translation model. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 523–530. Association for Computational Linguistics.

[121] Yamada, K. and Knight, K. (2002). A decoder for syntax-based statistical mt. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 303–310. Association for Computational Linguistics.

[122] Yang, N., Liu, S., Li, M., Zhou, M., and Yu, N. (2013). Word alignment modeling with context dependent deep neural network. In *ACL (1)*, pages 166–175.

[123] Zang, S., Zhao, H., Wu, C., and Wang, R. (2015). A novel word reordering method for statistical machine translation. In *Fuzzy Systems and Knowledge Discovery (FSKD), 2015 12th International Conference on*, pages 843–848. IEEE.

[124] Zeman, D. (2010). Using tectomt as a preprocessing tool for phrase-based statistical machine translation. In *Proceedings of the 13th international conference on Text, speech and dialogue*, TSD'10, pages 216–223, Berlin, Heidelberg. Springer-Verlag.

[125] Zens, R., Matusov, E., and Ney, H. (2004). Improved word alignment using a symmetric lexicon model. In *Proceedings of the 20th international conference on Computational Linguistics*, page 36. Association for Computational Linguistics.

[126] Zhang, H. and Chiang, D. (2014). Kneser-ney smoothing on expected counts. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 765–774, Baltimore, Maryland. Association for Computational Linguistics.

[127] Zhang, W., Yoshida, T., Tang, X., and Ho, T.-B. (2009). Improving effectiveness of mutual information for substantival multiword expression extraction. *Expert Syst. Appl.*, 36(8):10919–10930.

[128] Zhang, Y., Wu, K., Gao, J., and Vines, P. (2006). Automatic acquisition of chinese–english parallel corpus from the web. In *Advances in Information Retrieval*, pages 420–431. Springer.

[129] Zollmann, A. and Venugopal, A. (2006). Syntax augmented machine translation via chart parsing. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 138–141. Association for Computational Linguistics.